universität wien

DISSERTATION

Titel der Dissertation

Cross-Age Peer Tutoring in Physik

Evaluation einer Unterrichtsmethode

Verfasserin

Mag. Marianne Korner

angestrebter akademischer Grad

Doktorin der Naturwissenschaft (Dr. rer. nat.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 791 411
Dissertationsgebiet lt. Studienblatt:	Physik
Betreuer:	Univ.-Prof. Dr. Martin Hopf

Inhaltsverzeichnis

1. Einleitung

Unmittelbaren Anlass zu dieser Studie gaben die Testergebnisse österreichischer Schüler/innen bei PISA 2009 (OECD, 2010), und hier wiederum das schlechte Abschneiden der Schüler/innen der Sekundarstufe 1, speziell im naturwissenschaftlichen Bereich. Da sich traditioneller Unterricht als wenig wirksam herausstellt (Duit & Treagust, 2012), schien es notwendig, über alternative Formen des Unterrichts und über geeignete Lernumgebungen nachzudenken und eine eventuelle Wirksamkeit empirisch zu belegen. Cross-Age Peer Tutoring bietet sich nach allem, was in der Vergangenheit dazu geforscht wurde, dafür an.

Peer Tutoring ist eine Arbeitsform, bei der Lernende einander beim Lernen unterstützen, unabhängig von ihrem Alter, der Jahrgangsstufe, einer speziellen Ausbildung oder einem Thema. Innerhalb der Schule stellt es eine vielversprechende Unterrichtsform dar, wenn man den Fokus auf die Umsetzung konstruktivistischer Lerntheorien innerhalb geeigneter Lernumgebungen richtet. Diese sollen im besonderen Maße dazu geeignet sein, einen Konzeptwechsel der Lernenden von vorunterrichtlichen Vorstellungen hin zu wissenschaftlich anschlussfähigen Vorstellungen zu unterstützen (Widodo & Duit, 2004).

Peer Tutoring wurde in der Vergangenheit bereits im Zusammenhang zu einigen Kontexten untersucht. Dazu zählten die Verbesserung der Lesefähigkeit oder Mathematik-Nachhilfe (P. A. Cohen, Kulik, & Kulik, 1982; Robinson, Schofield, & Steers-Wentzell, 2005) genauso wie eine Schulung der Fertigkeiten am Computer (Fogarty & Wang, 1982) oder des Denkens (Topping, Peter, Stephen, & Whale, 2004). Darüber hinaus gibt es Hinweise, dass Peer Tutoring auch zu konzeptuellen Erfolgen im Zusammenhang mit naturwissenschaftlichem Unterricht führt (Howe, Tolmie, Greer, & Mackenzie, 1995; Lumpe & Staver, 1995). In seiner großen Meta-Meta-Studie attestiert Hattie (2009) dem Peer Tutoring starke Effekte mit Effektstärken von d = 0,55, die sogar noch gesteigert werden können, wenn Tutoren ein wenig älter sind als jene Schüler/innen, mit denen sie lernen. Man spricht dann von Cross-Age Peer Tutoring.

Die oben zitierten Studien berichten von positiven Effekten des Peer Tutorings auf die Schüler/innen, die sich nicht nur auf die wissensmäßige, kognitive Ebene beschränken,

sondern die auch positiv auf die Einstellungen zum Lernen und zur Schule wirken. Interessanter Weise berichten einige der Studien davon, dass diese positiven Effekte für alle Schüler/innen beobachtbar sind, egal ob sie aktiv als Tutoren am Prozess teilnehmen oder passiv, als die belehrten Schüler/innen, die als Tutees bezeichnet werden. Dabei wird davon ausgegangen, dass die Wirksamkeit des Peer Tutoring Prozesses auf der sozialen Interaktion der Schüler/innen beruht, da die Tutoren keine ausgebildeten Lehrer/innen sind (Fogarty & Wang, 1982; Lumpe & Staver, 1995). Diese Idee geht bereits auf Vygotsky zurück, der davon ausgeht, dass jede individuelle Entwicklung in der Gesellschaft und in der Kultur verwurzelt[1] ist (Vygotsky, 1978).

Das Ziel der hier vorgestellten Studie ist es zu untersuchen, ob Peer Tutoring als Unterrichtsmethode geeignet ist, die Qualität des Physikunterrichts zu verbessern und dies anhand von Evidenzen zu belegen. Ob sich die bereits bekannten Ergebnisse vorangegangener Studien auf den Physikunterricht übertragen lassen ist a priori nicht klar. Naturwissenschaftlicher Unterricht, insbesondere wenn er sich an konstruktivistischen Instruktionsansätzen orientiert, hat nämlich die entscheidende Rolle der vorunterrichtlichen Vorstellungen der Schüler/innen zu berücksichtigen und die Tatsache, dass diese im Zuge eines Prozesses, der als *Conceptual Change* bezeichnet wird, erst nach und nach zu wissenschaftlich anschlussfähigen entwickelt werden (Duit & Treagust, 2003). Die bisherige Forschung ergab, dass diese vorunterrichtlichen Vorstellungen sehr stabil sind (Duit, Treagust, & Widodo, 2008). Daher bedarf es geeigneter Lernumgebungen, die den Schüler/innen Anlass und Raum zur Reflexion und Adaption ihrer mitgebrachten Vorstellungen geben. Um genau an diese anschließen zu können, orientiert sich die vorliegende Studie an konkreten Themen und innerhalb dieser Themen an bereits bekannten Schülervorstellungen.

Bei jedem Lernprozess spielt neben der kognitiven, die emotionale Komponente eine wesentliche Rolle. Deswegen, und da Physikunterricht traditionell als uninteressant (Muckenfuß, 1995) empfunden wird, wurde entschieden, im Rahmen der Untersuchungen nicht nur die Entwicklung des Wissens in Richtung eines *Conceptual Change* zu untersuchen, sondern diese auch mit der Motivation der Schüler/innen in

[1] „... the mechanism of individual developmental change is rooted in society and culture“ (Vygotsky, 1978, S. 7)

Verbindung zu setzen. Um die Entwicklung des Wissens und seine Persistenz abbilden zu können, wurden vor und zu zwei Zeitpunkten nach der Peer Tutoring Intervention Wissenstests durchgeführt (Praetest – Posttest – Follow-up Test).

Diese Arbeit fokussiert somit vom Alter der Schüler/innen her auf die Sekundarstufe 1, vom Thema her auf spezielle Kontexte aus der Optik und der Elektrizitätslehre. Cross-Age Peer Tutoring wird als konstruktivistische Lernumgebung vorgestellt, um Physik zu unterrichten. Die Wirksamkeit dieser Methode soll basierend auf Wissenstests und motivationalen Parametern evaluiert werden.

2. Theoretischer Hintergrund

2.1. Konstruktivistische Lerntheorien und *Conceptual Change*

Konstruktivistische Sichtweisen bilden etwa seit den 1980er-Jahren einen unverzichtbaren theoretischen Rahmen für die Lehr- und Lernforschung, allen voran für den naturwissenschaftlichen Bereich.

Die altbekannte, immer wiederkehrende Phrase: „Aber ich habe es ihnen doch erklärt!“ beschreibt dabei treffend das Unverständnis von Seiten der Lehrkräfte, das den Ausgangspunkt konstruktivistischer Überlegungen bildet. Kognitivistische Modelle des Lehrens und Lernens gehen davon aus, dass ein Vermitteln von Information genügt, um einen Lernprozess in Gang zu setzen (Riemeier, 2012). Der Lehrende transportiert sein Wissen zum Lernenden, gleich einem Nürnberger Trichter soll es in die Köpfe der Lernenden eingetrichtert werden. Doch oft verstehen die Schüler/innen die vermeintlich klar kommunizierte Botschaft nicht. Aus diesem Grund ist eine Theorie des Verstehens (Häußler, Bünder, Duit, Gräber, & Mayer, 1998) gefordert. Diese Theorie verspricht konstruktivistische Ansätze zu bieten. In der Praxis zeigt sich, dass Unterrichtskonzepte auf Basis konstruktivistischer Lerntheorien auch besser als traditionelle Unterrichtskonzepte auf Basis kognitivistischer Modelle funktionieren (Duit & Treagust, 2012), jedoch nur unter der Voraussetzung, dass an bestehende Schülervorstellungen angeknüpft wird (Widodo & Duit, 2004). Daher ist es angebracht, sich dem Konstruktivismus im Zusammenhang mit der Lehr- und Lernforschung zu widmen.

Genau so wenig, wie es *den* Konstruktivismus gibt, gibt es in der Lehr- und Lernforschung die *eine* konstruktivistische Sichtweise. Vielmehr werden hier unterschiedliche Varianten des Konstruktivismus bemüht, die man zusammenfassend als moderaten oder neuen Konstruktivismus bezeichnen kann (Gerstenmaier & Mandl, 1995). Der moderate Konstruktivismus hat seine Wurzeln im sozialen Konstruktivismus und im radikalen Konstruktivismus, obwohl er mit letzterem weder Untersuchungsgegenstand noch Ansichten über Realität teilt.

2.1.1. Konstruktivistische Sichtweisen

Historisch gesehen müsste man den sozialen Konstruktivismus nach Berger und Luckmann an erster Stelle nennen (Knorr-Cetina, 1989). Diese ontologische Ausprägung des Konstruktivismus untersucht das Zustandekommen und die Aufrechterhaltung sozialer Ordnungen, oft auch gesellschaftlicher Machtstrukturen. Die zentrale Frage von Berger und Luckmann ist, wie es geschehen kann, dass vom Menschen selbstproduzierte soziale Ordnungen als „objektiv“ und „naturgegeben“ erfahren werden. Der Sozialkonstruktivismus beantwortet diese Frage, indem er drei ablaufende Prozesse nennt: An erster Stelle wird die Institutionalisierung, z.B. sozialer Rollen genannt. Parallel dazu läuft ein Objektivierungsprozess ab, der durch die Sprache gesteuert ist, wobei wiedergegebene Erfahrungen erzählt und dabei sprachlich typisiert werden. An dritter Stelle stehen Legitimationsprozesse, die als Objektivierungen zweiter Ordnung gesehen werden, indem sie die oben genannten Institutionalisierungsprozesse an jene vermitteln, die selbst nicht Teil dieser waren. Im Wesentlichen geht es hier um Vermittlung von Erfahrung und die Weitergabe von bereits ausgehandelten Bedeutungen. Als Legitimationen zweiter Ordnung wird das Formulieren von Theorien beschrieben, die auf einer Metaebene erklären, rechtfertigen und objektivieren (Knorr-Cetina, 1989).

Diese Sichtweise des sozialen Konstruktivismus ist insofern für die Fachdidaktik interessant, da sowohl die Sprache in einer Wissenschaft, als auch die wissenschaftlichen Theorien selbst als Ergebnis eines Aushandlungsprozesses gesehen werden, der auf Basis einer gesellschaftlichen Ordnung stattfindet. Die Physik bildet da keine Ausnahme. Jedenfalls legt bereits der soziale Konstruktivismus den Schluss nahe, dass auch naturwissenschaftliche Theorien im Kontext der jeweiligen Kultur und gesellschaftlichen Ordnung gesehen und interpretiert werden müssen, da sie das Ergebnis sozialer Aushandlungsprozesse sind und von Menschen erschaffen werden (Ertl, 2014).

Der erkenntnistheoretische, radikale Konstruktivismus versteht sich als eine Theorie des Wissens und beschäftigt sich mit dem Verhältnis von Wissen zur Welt. Er hat seine Wurzeln einerseits in der Neurobiologie, die das Gehirn lange Zeit als ein autopoietisches System auffasste (wie z.B. Maturana und Varela in ihren Arbeiten),

andererseits in der Philosophie von Kant und Wittgenstein, sowie in der Kognitionspsychologie nach Piaget (Knorr-Cetina, 1989). Kant schreibt in einem Vorwort zur „Kritik der reinen Vernunft“: *„Die Wirklichkeit [...] ist also durchwegs so, wie die menschliche Vernunft sie konstruieren kann.“* (in: Glasersfeld, 1997, S 10). Damit wird ausgeschlossen, dass der Mensch die Wirklichkeit erfassen kann, er kann sich lediglich ein Bild von ihr machen. Beobachtung wird im radikalen Konstruktivismus als Unterscheidung definiert, Wissen und Erkenntnis korrespondieren nicht mit der Welt. Wissen bleibt zwar immer empirisches Wissen, es ist aber intersubjektiv und daher vermittelbar. Es wird angehäuft, um als eine Orientierungshilfe in der Welt zu dienen. Das Kriterium für Wissen ist demnach nicht „wahr“ oder „falsch“[2], weil wir das gar nicht feststellen können, sondern die Viabiliät des Wissens, mit dem Ziel uns Orientierung in der Welt zu verschaffen (Knorr-Cetina, 1989).

Immerhin, es gibt im erkenntnistheoretischen Konstruktivismus einen Beobachter. Wahrnehmung funktioniert als Interpretation und Bedeutungszuweisung des Gehirns, das sich auch selbst beobachten kann. Somit besitzt es die Fähigkeit zur Reflexion. Diese Sicht der Wahrnehmung heißt aber, dass die Umwelt zwar Bedeutungen auslösen, aber nicht determinieren kann. Diese Sichtweise scheint für konstruktivistische Lerntheorien insofern von Bedeutung, als dass Lernumgebungen und Lernangebote zwar die Konstruktion von Wissen ermöglichen sollen, jedoch von einer Entwicklung in eine bestimmte, erwünschte Richtung nicht sicher ausgegangen werden kann.

In der Spielart des empirischen, radikalen Konstruktivismus beginnt man sich mit wissenschaftstheoretischen Fragestellungen zu beschäftigen. Er stellt sich prozessorientiert dar, und der eigentliche Untersuchungsgegenstand ist die Suche nach den Mechanismen der Welterschließung. Das bedeutet den Übergang von WARUM-Fragen zu WIE-Fragen. Der empirische Konstruktivismus beansprucht für sich, auf alle Bereiche anwendbar zu sein. Diese Forderung ist auch als Symmetriepostulat bekannt. Für die Naturwissenschaften und ihre Didaktiken bedeutet das, dass auch im Bezug auf naturwissenschaftliche Theorien von einer sozialen Konstruiertheit gesprochen wird. Daher kann in diesem Bereich der soziale und politische Hintergrund des

[2] Im Gegensatz zum kritischen Rationalismus von Popper, nach dem es auch keine „wahre“ oder „falsche“ Theorie gibt, leugnen die erkenntnistheoretischen Konstruktivisten die Existenz einer Realität, während für Poppers Überlegungen die Realität eine Voraussetzung ist.

Wissenschaftlers ebenso wenig ausgeklammert werden. Das ist ein Aspekt, der später im moderaten Konstruktivismus aufgegriffen wird. Das Kriterium von Wissen ist für den empirischen Konstruktivismus die Erweiterung unseres Aktionsradius in der Welt (Knorr-Cetina, 1989). Wie in der modernen Fachdidaktik, wird bereits hier die Nähe zum Untersuchungsgegenstand gefordert, was man in den konstruktivistischen Lerntheorien als Kontextorientiertheit wieder findet.

Eine der ersten Fachdidaktiken, in die konstruktivistische Überlegungen Eingang gefunden haben, ist die Physikdidaktik (Gerstenmaier & Mandl, 1995). Als Anfangspunkt kann der Artikel „Kinder konstruieren Welten“ von Aufschnaiter, Fischer und Schwedes gesehen werden (1992), die insofern in der Tradition des radikalen Konstruktivismus argumentieren, da sie von der semantischen Abgeschlossenheit und der Selbstreferentialität des autopoietischen Systems Gehirn ausgehen. Radikal auch deshalb, weil sie ihre Theorie selbst als Konstruktion der Konstruktion der Realität durch Schüler/innen sehen. Einen Kernpunkt dieser Theorie bildet die Tatsache, dass Schüler/innen nicht als unbeschriebene Blätter in den Unterricht kommen, sondern bereits über konkrete Vorstellungen zu vielen Bereichen verfügen (Alltagsvorstellungen). Wahrnehmung funktioniert radikal-konstruktivistisch gesehen so, dass das kognitiv und semantisch abgeschlossene Gehirn ankommenden Reizen auf Basis dieser Vorerfahrungen eine Bedeutung zuweist. Aus dieser Sichtweise lässt es sich klar argumentieren, dass Lernen *„... nicht das Ergebnis eines Einfüllvorganges von außen sein...[kann]*“ (Aufschnaiter, et al., 1992, S 386). Das stellt eine diametral entgegengesetzte Sicht zu kognitivistischen Lerntheorien dar und zeigt gleichzeitig eine mögliche Ursache dafür, dass kognitivistische Instruktionsansätze schlecht funktionieren. Lernen bedeutet in dieser radikal konstruktivistischen Sicht, dass sogenannte subjektive Erfahrungsbereiche weiter entwickelt werden. Neuen Problemen wird auf Basis der subjektiven Erfahrungsbereiche, die der Lerner bereits hat, nicht nur Bedeutung gegeben, sie werden gelöst. „Lösen“ bedeutet hier eine möglichst widerspruchsarme Konstruktion der neuen Situation in Verbindung mit den eigenen Vorerfahrungen. Das konkrete Tun dient dazu, Widersprüche, die wir bei der Betrachtung der Welt erfahren, zu minimieren. Sind die Widersprüche zur eigenen Konstruktion zu groß, ist es nicht mehr möglich, einen bereits existierenden subjektiven

Erfahrungsbereich zu adaptieren und es muss ein neuer geschaffen werden. Diese Prozesse entsprechen Piagets Begriffen von Akkommodation und Adaption (vgl. S. 18). Die Rolle der Fachdidaktik besteht nun darin, die dafür geeigneten Lernumgebungen zu schaffen und die Entwicklung der subjektiven Erfahrungsbereiche zu dokumentieren und zu unterstützen.

2.1.2. Moderater Konstruktivismus

Im moderaten Konstruktivismus geht man wie im radikalen Konstruktivismus davon aus, dass alles Wissen menschlich konstruiert ist. Jedoch wird eine solipsistische Sicht vermieden und die Existenz einer „Realität", hauptsächlich aus pragmatischen Gründen, nicht infrage gestellt (Widodo & Duit, 2004). Der radikale Konstruktivismus dreht sich um die Frage, was Erkenntnis ist und wie sie entsteht. Der moderate Konstruktivismus hingegen ist eine Auffassung vom Lernen. Es wird der Fokus auf die Frage gelegt, *wie* sich Erkenntnisse verändern und, im Rahmen konstruktivistischer Lerntheorien, individuell verändern lassen (Riemeier, 2012).

Das führte in der Vergangenheit zu einer Reihe konstruktivistischer Instruktionsansätze. Allen gemeinsam ist, dass die Lernenden im Mittelpunkt steht und Lernen als Konzeptwechsel (*conceptual change*), bzw. als Konzeptentwicklung (*conceptual growth*) gesehen wird (Häußler, et al., 1998).

Zu den drei wichtigsten konstruktivistischen Instruktionsansätzen gehört der *Anchored Instruction* Ansatz nach Bransford und Franks. Ausgangspunkt der Überlegungen ist die Fragestellung, wie verhindert werden kann, dass träges Wissen entsteht. Darunter versteht man ist eine Art von Wissen, das zwar reproduziert, aber nicht angewandt werden kann, weil es nicht mit dem Vorwissen in Verbindung gebracht werden kann und daher zusammenhangslos ist. Mit narrativen Ankern will man zunächst Interesse erzeugen. Ein wichtiger Punkt dabei ist die Dekontextualisierung von Wissen, daher die Fertigkeit, Wissen in unterschiedlichen Kontexten anwenden zu können, indem man es von Vornherein auf unterschiedliche Art anwendet. Leider existieren zu diesem Ansatz wenige empirische Untersuchungen (Gerstenmaier & Mandl, 1995).

Ein weiterer Ansatz ist der *Cognitive Flexibility* Ansatz nach Spiro et al. In ihrer *Random Access Theory* plädieren sie dafür, dass multiple Perspektiven eingenommen werden

sollen, daher Problemstellungen aus verschiedenen Blickwinkeln betrachtet werden sollen (Gerstenmaier & Mandl, 1995; Widodo & Duit, 2004). Die Autoren Gerstenmaier und Mandl geben an, dass diese Art der Instruktion auf wenig strukturierten Gebieten und für fortgeschrittene Lerner geeignet ist. Auch zu diesem Ansatz liegen bezüglich seiner Wirksamkeit kaum empirische Untersuchungen vor.

Jener konstruktivistische Instruktionsansatz, der im Zusammenhang mit der Untersuchung, die in dieser Arbeit beschrieben wird, am interessantesten ist, ist der *Cognitive Apprenticeship* (CAS) Ansatz nach Collins (Collins, Brown, & Holum, 1991; Collins, Brown, & Newman, 1989), das kognitive Lehrlingsmodell. Collins et al. entwickelten es nach Vygotskys Modell des sozialen Lernens: Vygotsky, der in der Tradition von Marx und Engels seine Theorien entwickelte, sah die Ursache in jeder individuellen Entwicklung in Kultur und Gesellschaft, also im Sozialen, verwurzelt (Vygotsky, 1978). Collins et al. übertrugen diese Idee auf die Entwicklung in der kognitiven Ebene.

Der Ansatz von Collins unterscheidet zwischen leicht zugänglichem Gegenstandswissen und dem lediglich implizit vorhandenen Expertenwissen. Um auch dieses vermitteln und nicht-sichtbare kognitive Vorgänge vermitteln zu können schufen sie in Anlehnung an die traditionelle Handwerkslehre ein vierstufiges Konzept: *modeling* (modellieren) – *coaching* (angeleitetes Üben) – *scaffolding* (Lerngerüst geben) und *fading-out* (sich zurückziehen). In der ersten Phase wird eine Problemlösungsstrategie in einer authentischen Situation von Experten modelliert und verbalisiert (lautes Denken). Damit enthält dieser Ansatz zusätzlich zu den beiden anderen, oben berichteten Ansätzen, ein Element der expliziten Anleitung. In der zweiten Phase lösen die Lernenden unter Anleitung der Experten mehr oder weniger selbstständig ein eigenes Problem und werden dazu angehalten die Lösungsstrategien zu verbalisieren. Mit wachsendem Lernfortschritt wird die Unterstützung der Experten weniger (scaffolding), bis sie schließlich gänzlich verebbt (fading-out). Die darüber hinaus zentralen Aspekte des CAS beinhalten eine Situiertheit des Lernens, das den Alltagsnutzen des Gelernten sichtbar machen soll und es wird der soziale Aspekt des Lernens betont. Durch unterschiedliche Problemstellungen und Kontexte soll eine Dekontextualisierung des Wissens erreicht werden.

Im Unterschied zu den Ansätzen *Anchored Instruction* und *Cognitive Flexibility* existieren zur Lernwirksamkeit des CAS einige empirische Untersuchungen, die der Methode hinsichtlich des Lösens von Anwendungsaufgaben insgesamt große Lernzuwächse bescheinigen (Gerstenmaier & Mandl, 1995, S. 879).

2.1.3. Charakteristik konstruktivistischer Instruktionsansätze

Will man konstruktivistische Instruktionsansätze charakterisieren, so fallen einige gemeinsame Merkmale auf, die mehr oder weniger explizit unterstützt werden und die sich aus den unterschiedlichen konstruktivistischen Strömungen entwickelt haben (Gerstenmaier & Mandl, 1995; Riemeier, 2012)

- Lernprozesse finden in authentischen Situationen statt (Authentizität und Situiertheit des Lernens).
- Den Lernenden soll die Gelegenheit gegeben werden in multiplen Kontexten multiple Perspektiven einzunehmen.
- Lernen erfolgt immer im sozialen Kontext, wobei kooperative Lernformen einen zentraler Punkt aller Ansätze darstellen.
- Lernen erfolgt immer individuell auf Basis der persönlichen (Vor-) Erfahrungen der Lernenden.
- Lernen erfolgt in speziell gestalteten Lernumgebungen, die zu einer Konstruktion des Wissens anregen.
- Worauf diese Lernprozesse hinsteuern, ist von außen nicht determinierbar oder kontrollierbar. Lernprozesse können lediglich initiiert und unterstützt werden.

Der letzte Punkt stellt nun die zentrale Fragen an die Fachdidaktik dar: Wie lassen sich konstruktivistische Prozesse initiieren und unterstützen? Welche Rolle nehmen Lehrende und Lernende in einem konstruktivistischen Unterricht ein? Wie können Lernumgebungen aussehen, die jedenfalls imstande sein sollen, die Schüler/innen zu fachlich adäquaten Konstruktionen anzuregen.

Unabdingbar ist , dass sich die Rolle der Lehrenden somit von der reinen Fachperson für den Stoff hin zu einer Fachperson für die Inhalte *und* das Lernen wandelt (Reusser, 2001). An die Stelle eines Wissensvermittlers, der seine „Ware“ darstellt und zu

verkaufen sucht, tritt ein Coach und Lernberater, der durch sein Verhalten selbst ein kognitives *rolemodel* im Sinne des CAS für seine Schüler/innen darstellt.

Widodo und Duit (2004) und Widodo (2004) bilden auf Basis einer Literaturrecherche die Charakteristika konstruktivistischer Lernumgebungen im Kategoriensystem KONU[3] (Konstruktivistisch Orientierter Naturwissenschaftlicher Unterricht) ab. Die Untersuchungen fanden im Rahmen einer Studie über Unterrichtsstunden in Deutschland statt, um herauszufinden, inwieweit konstruktivistische Ansätze in den „gewöhnlichen" Unterricht einfließen.
Fünf Eigenschaften charakterisieren nach Widodo und Duit (ebd.) konstruktivistische Lernumgebungen: Sie

- ermöglichen die Konstruktion von Wissen, indem bedeutungsvolle, authentische Probleme angeboten werden,
- berücksichtigen vergangene (Lern-) Erfahrungen, das Vorwissen, die Interessen und die Lernbedürfnisse,
- geben Raum zur sozialen Interaktion, weil Wissen sozial konstruiert ist,
- unterstützen die Schüler/innen beim eigenständigen Lernen und stärken die Selbstverantwortung und
- berücksichtigen die konstruktivistische Sicht von Wissenschaft, dass sich Wissen und Theorien weiter entwickeln (*Nature of Science*).

Will eine Fachdidaktik die Qualität einer Lernumgebung beschreiben, so ist jedenfalls die Passung wichtig: Problemstellungen müssen so formuliert werden, dass sie auf Basis der Vorerfahrungen zum Konstruieren neuer Vorstellungen anregen. Darüber hinaus ist zu beantworten, wie viel direkte Instruktion benötigt wird, um einen zielgerichteten Instruktionsprozess zu initiieren (Riemeier, 2012). Einen vorläufigen Schlussstrich unter diese Debatte versuchen Clark, Kirschner und Sweller (2012) zu ziehen, indem sie eine *minimally guided instruction* einer *direct instruction* gegenüberstellen. Es scheint sinnvoll zu sein, dass Lehrer/innen zu Beginn explizite Instruktion betreiben sollen, um sich danach in dem Maße zurückzuziehen, in dem das Wissen der Lernenden zunimmt. Es wird argumentiert, dass ein Zurückhalten von Information keine Konstruktionsprozesse fördert. Diese Sichtweise ist durchaus mit der Idee des CAS zu vereinbaren.

[3] Bzw. COSC – Constructivist Oriented Science Classroom.

Gefordert wird von einem Unterricht auf Basis konstruktivistischer Lerntheorien darüber hinaus, dass er sich an den Lernenden orientiert und nicht bloß am Fach. Lehrpersonen müssen auch die Perspektive der Schüler/innen einnehmen können und bereit sein, daraus selbst zu lernen (Riemeier, 2012).

Widodo und Duit (2004) merken allerdings an, dass sich die Didaktik der Naturwissenschaften nicht auf die Entwicklung von Lernumgebungen beschränken soll, sondern auch Interessen, Bedürfnisse und kognitive Fähigkeiten der Schüler/innen berücksichtigen muss.

Um Aussagen über die Mechanismen der Konzept*entwicklung* der Schüler/innen machen zu können, bedarf es mehr als einer konstruktivistischen Lerntheorie. Sie ergänzend lassen sich die *Conceptual Change* Theorie und die Ergebnisse der seit Jahrzehnten existierenden Forschung und Katalogisierung von Schülervorstellungen in die Konstrukte kognitivistischer Lerntheorien einpassen, um Aussagen über Konzeptentwicklungen treffen zu können (Riemeier, 2012).

2.1.4. *Conceptual Change* Theorien

Conceptual Change Theorien bilden innerhalb konstruktivistischer Unterrichtsansätze das theoretische Gerüst dafür, welche Vorstellungen, die Schüler/innen bereits in den Unterricht mitbringen, sie im Rahmen von Lernprozessen verändern, wie diese Veränderung abläuft und unter welchen Bedingungen es überhaupt dazu kommt. Mittlerweile bildet dieses theoretische Gerüst die führende Basis für Lehr- und Lernforschung (Duit & Treagust, 2012).

Bereits Ausubel stellte 1968 fest: *„The most important single factor influencing learning is what the learner already knows. Ascertain this and teach ... accordingly“* (in: Widodo & Duit, 2004, S 235). Das kann als ein Ausgangspunkt der *Conceptual Change* Theorie gesehen werden, wobei in den letzten Jahrzehnten viel Forschung dahingehend investiert wurde, was *„teach accordingly“* bedeutet und welche Möglichkeiten der Umsetzung es gibt.

Der Begriff des *Conceptual Change* wurde einerseits auf Basis von Thomas Kuhns Konzept des Paradigmenwechsels in der Wissenschaft gebildet und geht andererseits auf Piagets Begriffe der Assimilation und der Akkommodation zurück (Krüger, 2012;

Vosniadou, 2013). Die Assimilation bezeichnet einen Vorgang, bei dem neue Erfahrungen auf Basis bereits vorhandener kognitiver Strukturen erklärt werden. Falls diese Erfahrungen mit Hilfe der alten Strukturen nicht mehr erklärt werden können, müssen auch die kognitiven Strukturen adaptiert werden und man spricht von der Akkommodation.

Posner, Strike, Hewson und Gertzog (1982) arbeiteten zum *Conceptual Change* eine Umsetzung aus, die als klassischer *Conceptual Change* Ansatz bekannt ist. In dieser Auffassung sind die mitgebrachten Vorstellungen der Lernenden inkorrekt, müssen aufgegeben werden und durch wissenschaftlich korrekte Vorstellungen ersetzt werden. Es müssen dabei vier Bedingungen erfüllt sein, damit Schüler/innen die Notwendigkeit eines Konzeptwechsels erkennen und bereit sind, sich darauf einzulassen: Sie müssen zunächst mit ihren bestehenden Konzepten in der aktuellen Lernsituation unzufrieden sein. Das neue Konzept, das im Rahmen des Unterrichts angeboten wird, und das möglichst übernommen werden soll, muss verständlich sein (*intelligible*), plausibel, in dem Sinne, dass eine gewisse Übereinstimmung zu den alten Konzepten und epistemologischen Überzeugungen existieren muss, es daher glaubhaft ist. Schließlich muss das neue Konzept erfolgreich anwendbar sein (*fruitful*), nicht nur in der aktuellen Situation, sondern auch für folgende. Es ist wichtig, dass auch die Lerner in der aktuellen Lernsituation diese vier Bedingungen wahrnehmen.

Diese vier Bedingungen beziehen sich allerdings auf rein kognitive Betrachtungen. Daher ist der klassische *Conceptual Change* Ansatz nicht unumstritten, zumal er keine affektiven Parameter wie Motivation oder Interesse berücksichtigt, die den Lernprozess mitsteuern (Duit & Treagust, 2012; Vosniadou, 2013). Aufgrund der Forderung nach einer Einbindung affektiver Haltungen (z.B. Zembylas, 2005) wurde ein multi-perspektivischer Ansatz entwickelt. Es werden dabei *Conceptual Changes* nicht nur auf der Inhaltsebene betrachtet, sondern auch auf Metaebenen. Dazu gehören die Ebene der *Nature of Science* und die Ebene der metakognitiven Einstellungen zum Lernen (Duit & Treagust, 2003). Allerdings gibt es hier noch offene Fragen zum theoretischen Rahmen, vor allem was die Interaktion zwischen diesen *Conceptual Change*s betrifft. Aus diesem Grund ist ein multi-perspektivischer *Conceptual Change* Ansatz in der Schulpraxis noch schwerer umsetzbar als der klassische (Duit & Treagust, 2012).

Der Begriff des Conceptual Change, des Konzeptwechsels, darf nicht als ein Auswechseln der naiven oder wissenschaftlich nicht anschlussfähigen Vorstellungen, die Schüler/innen haben und in den Unterricht mitbringen, hin zu wissenschaftlich korrekten Vorstellungen verstanden werden. Er ist vielmehr als eine Weiterentwicklung der Konzepte zu verstehen, was sich auch in unterschiedlichen Sprechweisen wie z.B. *conceptual growth* oder *conceptual development* zeigt. Ausgehend von der konstruktivistischen Sichtweise, dass die Basis jeder Neukonstruktion immer die bereits bestehenden Konzepte sind, kann man bei *Conceptual Change* nicht von einem Austausch der Konzepte ausgehen. Die bereits bestehenden Konzepte werden niemals völlig aufgegeben. In manchen Kontexten werden daher auch die alten Konzepte wieder zur Anwendung kommen, wenn sie viabler erscheinen (Duit & Treagust, 2003). Außerdem ist die Konzeptentwicklung eben davon abhängig, was an individuellen Konzepten bereits vorhanden war. Somit ist das Lernen im Sinne eines Konzeptwechsels immer individuell.

Eine weiter führende, für die fachdidaktische Forschung interessante Frage ist diejenige, wie ein *Conceptual Change* erreicht werden kann. Dieser kann einerseits auf sanfte Weise, durch Umknüpfen des Begriffsnetzes erreicht werden. Dabei kommt es zu einer graduellen Veränderung der Konzepte und nicht zu einem plötzlichen Umlernen. Andererseits kann der Begriffswechsel auf radikalere Weise, entsprechend Piagets Idee von Akkommodationsprozessen, durch einen kognitiven Konflikt erreicht werden. Er tritt immer dann auf, wenn die vorhandenen kognitiven Strukturen nicht in Einklang mit Beobachtungen oder Daten zu bringen sind, es also notwendig wird, neue kognitive Strukturen zu entwickeln. Im Rahmen von Unterricht können entsprechend dieser Vorstellung genau solche Lernsituationen geschaffen werden, die einen kognitiven Konflikt initiieren.

In letzter Zeit gibt es Strömungen in der Fachdidaktik, die zur ersteren, sanfteren Variante tendieren, da im Falle des kognitiven Konfliktes aus mehreren Gründen Vorsicht geboten ist: Es ist ein Konflikt, der von den Lehrenden bewusst herbeigeführt wird. In manchen Fällen wird der zu diskutierende Widerspruch von den Lernenden jedoch gar nicht wahrgenommen oder er erscheint ihnen irrelevant. Es kann auch passieren, dass er nur von einem Teil der Lernenden wahrgenommen wird. Eine

mögliche Erklärung durch die Lehrenden, wo denn der Konflikt eigentlich liegen soll und welche Bedeutung er haben soll, kann sich hierauf wiederum zu einem kognitivistischen Unterrichtsstil entwickeln. Außerdem kann es passieren, dass sich die Begriffsstrukturen der Lernenden nicht entwickeln, weil die Diskrepanz zu den bereits vorhandenen Begriffsstrukturen zu groß oder zu klein ist, jedenfalls keine Akkommodationsprozesse initiiert werden. Darüber hinaus ist es möglich, dass sich die Begriffsstrukturen nicht in die gewünschte Richtung entwickeln (Duit & Treagust, 2012; Krüger, 2012).

Für Forschungsdesigns ergibt sich zusammenfassend die Konsequenz, dass neben einer inhaltlichen Konzeptentwicklung auch affektive Komponenten berücksichtigt werden sollten. Für die Gestaltung von Lernumgebungen sind die oben genannten Kriterien zu erfüllen, insbesondere, dass sie variable, für die Schüler/innen persönlich bedeutsame Kontexte beinhalten und dass sie über eine entsprechende Passung verfügen, was die zu erzeugende Unzufriedenheit und die Verständlichkeit der alternativ angebotenen Konzepte betrifft. Hinsichtlich einer Rahmentheorie zur *Nature of Science* müssen metakognitive Fähigkeiten der Lernenden gefördert werden, damit wissenschaftliche Konzepte für Schüler/innen glaubwürdiger werden und somit besser angenommen werden können.

Betreffend die Forschung zum *Conceptual Change* im Zusammenhang mit kollaborativen Lernformen gibt es den positiven Befund, dass die Zusammenarbeit mit anderen die erwünschte Quelle unterschiedlicher Perspektiven auf ein Problem darstellt, was in weiterer Folge zu einer Konzeptentwicklung führen kann (Miyake, 2008). In ihrer Studie wird das *constructive interaction framework* vorgestellt. Dabei bekamen Schüler/innen die Anweisung *„talk together and figure out how .... works“* (ebd., S 464), was eine weniger gekünstelte Alternative zum Laut-Denken darstellen soll. Die Transskripte werden in weiterer Folge analysiert. Ausgegangen wird von der Annahme, dass die grundlegende Form des *Conceptual Change* mit einer Interaktion zwischen inneren und äußeren Ressourcen der Lernenden verbunden ist. Die inneren Ressourcen sind dabei die bereits angehäuften Strukturen (Wissen), die äußeren stellen die physische und soziale Umwelt der Lernenden dar. Im Rahmen der Interaktionen werden Lösungen sortiert und in bereits vorhandene Konzepte integriert, was als Quelle für einen *Conceptual Change* gesehen wird.

Eine Studie von Howe, Tolmie, Greer und Mackenzie (1995) beschäftigt sich mit dem Zusammenhang zwischen kollaborativen Lernformen, *Conceptual Change* und Physik. Es findet sich schon in Piagets Werk ein Hinweis darauf, dass die Zusammenarbeit zwischen Peers zu einer konzeptuellen Weiterentwicklung führen kann (ebd.), vorausgesetzt, dass sie über unterschiedliche Konzepte verfügen und dass die Aufgabenstellungen zur Auseinandersetzung damit anregen. Studien aus den 1980-er Jahren lieferten dazu vielversprechende Resultate. Dennoch, eine sorgfältige Evaluierung war hier nicht zu finden, was die Autoren dieser Studie anhand speziell designter Aufgaben in einem Prae-Posttestdesign nachholten. Die Autoren kommen zur Conclusio, dass Peer Tutoring im Zusammenhang mit dem untersuchten thermodynamischen Thema einen *Conceptual Change* anregen kann.

Insgesamt kann gefolgert werden, dass sich nach dem Studium der Literatur durchaus der Eindruck ergibt, dass, ausgehend von konstruktivistischen Auffassungen vom Lernen, kollaborative Lernformen, zu denen Peer Tutoring ebenfalls gezählter werden kann, geeignet sind, um einen *Conceptual Change* einzuleiten. Allerdings gibt es wenige Arbeiten, die anhand konkreter Schülervorstellungen aus Themenbereichen der Physik mögliche Erfolge des Peer Tutoring zeigen.

2.2. Peer Tutoring

Peer Tutoring ist keine Erfindung moderner Lehr- und Lernforschung, sondern kann bis hin zu den alten Griechen zurückverfolgt werden (Topping, 1996). Dennoch fügt es sich in die Reihe moderner, moderater konstruktivistischer Unterrichtsansätze. Es kann in der Form des *Cross-Age Peer Tutoring* (CAPT) als eine Spielart oder Weiterentwicklung des *Cognitive Apprenticeship* gesehen werden. Jedenfalls bietet es eine interessante Möglichkeit für Lernende, in ihren Lernprozess aktiv einzugreifen und ihn selbstbestimmt zu gestalten.

2.2.1. Historische Betrachtung

Historisch gesehen war und ist es in vielen Kulturen üblich, dass ältere Kinder ihre jüngeren Geschwister betreuen, für sie Verantwortung übernehmen und somit quasi in die Rolle von Tutoren schlüpfen. Dies kann durchaus als eine Form des Peer Tutoring

gesehen werden, allerdings als eine unorganisierte. Im Zeitalter der Industriellen Revolution wurden Peers, also Schulkollegen, gezielt eingesetzt, um den entstandenen Lehrermangel, vor allem in den Schulen der Arbeiterklasse, zu kompensieren (Fogarty & Wang, 1982). Ein weiteres großes Betätigungsfeld für Tutoren ist auch die klassische Nachhilfe oder das Wiederholen von Lerneinheiten in Tutorien an Colleges oder Universitäten, um Dropout-Raten zu senken. Tutoren wurden lange als Ersatzlehrer in einem mehr oder weniger linearen Modell der Wissensvermittlung gesehen (Topping, 1996).

In den letzten Jahrzehnten, als man begann, sich mehrheitlich konstruktivistischen Lerntheorien zu widmen, wurde Peer Tutoring Gegenstand bildungswissenschaftlicher Forschung (Robinson, et al., 2005). Damit wandelte sich aber auch die Auffassung über die Art und Qualität der Interaktionen zwischen Tutoren und deren Schützlingen, den Tutees. Tutoren werden heute nicht mehr als schlecht oder gar nicht ausgebildete, billige Ersatzlehrer gesehen, denn dann dürften sie wenige bis keine Unterrichtserfolge erzielen. Dazu geben aber die mehrheitlich guten Befunde zur Lernwirksamkeit des Tutorings keinen Anlass. Daher muss die Tutoren-Tutee-Interaktion anders funktionieren als die Lehrer-Schüler-Interaktion und auch die Wissensvermittlung dürfte über andere Mechanismen laufen (Fogarty & Wang, 1982; Topping, 1996).

2.2.2. Definition von Peer Tutoring und Cross-Age Peer Tutoring

Vor dem Hintergrund dieses Wandels im Bild der Tutoren benötigen wir eine moderne Definition, die sich in konstruktivistische Lerntheorien einbetten lässt und auf Basis derer die vielfältigen Aspekte der Tutor-Tutee-Interaktionen erforscht werden können. Gewählt wurde die folgende Definition von Topping (2005): *„[Peer Learning] involves people from similar social groupings who are not professional teachers helping each other to learn and learning themselves by so doing“* (S 631). Diese Definition ist sehr offen, weist aber bereits auf einen Wandel im Forschungsfokus neuerer Publikationen hin: weg von einer ausschließlichen Betrachtung der kognitiven und affektiven Entwicklung der Tutees hin zu jener der Tutoren. Die Begriffe Peer Learning oder, wenn die Rollenverteilung strikt geregelt ist, Peer Tutoring beziehen sich auf Lernende gleichen Alters, während Cross-Age Peer Tutoring unterschiedliche Altersstufen der Tutoren und Tutees bezeichnet.

Um den Begriff *Cross-Age* Peer Tutoring zu präzisieren, wurde auf Gaustad (1993) zurückgegriffen: *„Peer tutoring occurs when tutor and tutee are at the same age whereas in cross-age peer tutoring the tutor is older than the tutee“* (S 1). Offen bleibt hier, um wie viel älter die Tutoren sein sollen als ihre Tutees.

2.2.3. Empirische Befunde früherer Studien zu Cross-Age Peer Tutoring

Frühere Studien und Metastudien zu Peer oder Cross-Age Peer Tutoring beschreiben zahlreiche Themen und unterschiedliche Altersgruppen, wobei Tutoring-Programme hinsichtlich unterschiedlicher Gesichtspunkte analysiert wurden.

Effekte im Überblick

Zunächst sei festgehalten, dass zahlreiche Studien dem Peer Tutoring positive Effekte hinsichtlich kognitiver und affektiver Parameter bescheinigen. Will man sich einen ersten Eindruck hierzu verschaffen, eignet sich hierfür die Meta-Meta-Studie von Hattie (2009). Er berechnet für Peer Tutoring eine mittlere Effektstärke[4] von 0,55. Damit fällt es als Unterrichtsmethode in die von ihm so bezeichnete Zone der *desired effects*: Hattie (ebd.) legt diesen Bereich mit Effektstärken größer oder gleich 0,4 fest und argumentiert, dass man im Kontext von Unterricht in diesem Bereich von überdurchschnittlichen Verbesserungen sprechen kann (S. 16).

Cohen, Kulik und Kulik (1982) verglichen in einer Metastudie die Ergebnisse von 65 Einzelstudien, die unterschiedliche Fokusse verfolgten, ebenfalls über das Maß der Effektstärke miteinander. Eine Mehrheit dieser Studien bescheinigt dem Peer Tutoring größere Effekte als herkömmlichem Unterricht, was auch andere Studien zeigen (Robinson, et al., 2005). Cohen et al. (ebd.) geben eine mittlere Effektstärke von 0,4 an, die jedoch von der gewählten Art des Tutorings abhängt. So zeigen kürzere, strukturiertere Programme, die auf cross-age Basis arbeiten und herkömmlichen Unterricht ersetzen (also nicht zusätzlich ablaufen) bessere Erfolge. Das sind Folgerungen, die auch von anderen Studien unterstützt werden, z.B. berichten Robinson et al. (2005) ebenfalls davon, dass längere Programme nicht notwendigerweise größere Erfolge verzeichnen.

[4] Die Effektstärke entspricht dem Mittelwertunterschied zweier Gruppen dividiert durch die Standardabweichung (SD). Bei unterschiedlichen SDs wird mit der gepoolten SD gerechnet.

Effekte auf Tutoren

Die Wirkung von CAPT auf Tutoren wird erstmals bei Cohen beschrieben (1982). In 33 von 38 Studien, die zu diesem Thema analysiert wurden, wird den Tutoren bescheinigt, dass sie nach dem Tutoring in kognitiver Hinsicht bessere Erfolge als Kontrollgruppenschüler/innen erzielen. Die mittlere Effektgröße des Tutorings betrug hier 0,33. Auch verbesserten sich für Tutoren deren Einstellungen zum Lernen mit einer Effektstärke von 0,42. Für Tutees konnte eine Verbesserung in der Einstellung zum Lernen nicht in einem statistisch bedeutsamen Maße nachgewiesen werden. Einen positiven Effekt bezüglich der verbalen Fähigkeiten der Tutoren bestätigen auch Fogarty et al. (1982). Ergebnisse aktuellerer Studien fasst Toppings Artikel über „Trends in Peer Learning“ (2005) zusammen und betont dabei ebenfalls die positiven Effekte auf die Helfer, vor allem in kognitiver Hinsicht. In ihrer Besprechung der Literatur aus demselben Jahr zu Peer Tutoring weisen auch Robinson et al. (2005) darauf hin, dass Tutoren teilweise überraschende kognitive Fortschritte beim Tutoring machen, selbst wenn sie davor keine zusätzliche inhaltliche Instruktion bekommen haben. Auch Hattie (2009) kommt zum Schluss, dass Peer Tutoring beeindruckende Effekte im kognitiven, wie im nicht-kognitiven Bereich für Tutees *und* für Tutoren hervorbringt.

Was die Durchführung von Tutoring-Programmen betrifft so gibt es die Möglichkeit, die Tutoren zuvor einem speziellen Training zu unterziehen. Dieses Training wird im Rahmen der vorliegenden Arbeit als „Mentoring“ bezeichnet. Einige Publikationen (z.B. Fogarty & Wang, 1982; Robinson, et al., 2005) fordern ein derartiges Training explizit ein, während es in anderen unerwähnt bleibt. Speziell, wenn die zu lehrenden Inhalte neu sind oder neue Konzepte thematisieren, ist es demnach sinnvoll, ein Mentoring durchzuführen. Ginge es in Tutorings z.B. um die Verbesserung von Lesefertigkeiten, wäre ein zusätzliches Tutorentraining eventuell verzichtbar. Vor dem Hintergrund eines angestrebten Konzeptwechsels im Kontext der Physik und dem Wissen um Schülervorstellungen erscheint das Mentoring umso angebrachter.

Peer Tutoring auf Basis unterschiedlicher Inhalte

Analysiert man die Literatur zum Peer Tutoring, so sieht man, dass sich diese Unterrichtsmethode für eine Reihe von Themengebieten bereits bewährt hat. So wird im Rahmen einer Metastudie berichtet, dass sie im Zusammenhang mit Mathematik besser

funktioniert als beim Lesen oder relativ unklar gelassenen „anderen“ Bereichen (P. A. Cohen, et al., 1982). Auch die Besprechung der Literatur von Robinson et al. (2005) berichtet, dass die Inhalte der untersuchten Studien im Wesentlichen auf Mathematik beschränkt waren. Bereits Fogarty und Wang (1982) wenden in ihrer Einzelstudie Tutoring für Mathematiknachhilfe und zur Verbesserung der Fähigkeiten am Computer an. Hingegen nehmen Rohrbeck, Ginsburg-Block, Fantuzzo und Miller (2003) in ihrem metaanalytischen Überblick alle jene Studien auf, die mit Inhalten arbeiteten, die sie als „akademisch“ bezeichneten.

In dem Maße, in dem die Mathematik in Studien zum Peer-Learning überrepräsentiert erscheint, werden Themen aus den Naturwissenschaften und im Speziellen der Physik, spärlich behandelt. Zu erwähnen wäre eine Studie über *Peer Collaboration* in Physik (Howe, et al., 1995), wo Kinder über Heiz- und Kühlprozesse arbeiten. Die Autoren berichten, dass diese Methode einen Konzeptwechsel im Zusammenhang mit diesen speziellen Themen ermöglicht. Hier ergibt sich ein themenbezogener zusätzlicher Forschungsbedarf. Eine weitere Studie (Lumpe & Staver, 1995) berichtet über *Peer Collaboration* im Zusammenhang mit einem Thema aus der Biologie (Photosynthese). Die Autoren kommen zum Schluss, dass *Peer Collaboration* zu einem *Conceptual Change* führen kann. Sie betonen aber, dass *Peer Collaboration* etwas anderes ist als Peer Tutoring, da die Prinzipien der Gegenseitigkeit und der Gleichheit dort nicht befolgt werden.

Lernen durch Lehren

Eine weitere Arbeit aus jüngerer Zeit beschäftigt sich ebenfalls mit Themen aus der Naturwissenschaft (Zinn, 2008, 2009). Allerdings wird hier auf das theoretische Konzept des Lernen durch Lehren (LdL) nach Martin (1998) zurückgegriffen, das in gewisser Weise in der Durchführung mit dem Peer Tutoring verwandt ist. Es wurde entwickelt, um dem streng lehrerzentrierten, frontalen Sprachunterricht im behavioristischen Stil der 1980er-Jahre entgegenzutreten. LdL versteht sich durchaus im Sinne konstruktivistischer Unterrichtsansätze. Folgende Vorgangsweise wurde gewählt: Eine Gruppe von Schüler/innen bereitet neue Inhalte für ihre Mitschüler/innen vor, die sie ihnen dann anschließend vermitteln. Durch diese aktive Auseinandersetzung mit den Inhalten sollen das Wissen vertieft, die Motivation gesteigert und auch nicht-kognitive

Fähigkeiten erprobt werden. Außerdem, und das ist für den Sprachunterricht, für den es entwickelt wurde, von besonderer Bedeutung, sollen sich dadurch die Sprechanteile der Schüler/innen im Unterricht erhöhen. Die Wirksamkeit dieser Methode wurde im sprachtheoretischer Hinsicht von Martin selbst untersucht (Martin, 1985). Grzega und Schöner (2008) identifizierten in ihrer Arbeit über LdL Anforderungen an einen zeitgemäßen Unterricht, um in einer Informationsgesellschaft besser zurecht zu kommen. Ihrer Auffassung nach entspricht LdL diesen Anforderungen. Darüber hinaus gibt es außer der Pilotstudie von Zinn (2009) wenig Empirisches Werk zu LdL. Er arbeitete ebenfalls auf cross-age Basis, wobei Mittelschüler/innen mit Grundschul- und Vorschulkindern zusammenarbeiteten. Die Themen kamen zwar alle aus der Naturwissenschaft, jedoch arbeiteten unterschiedliche Gruppen zu unterschiedlichen inhaltlichen Fragestellungen daran. Zinn betont in seiner Conclusio, dass das Interesse der Schüler, speziell aber auch der Schülerinnen, durch LdL gestiegen ist und dass das eher implizit vorhandene Prozesswissen zur Anwendung kommt. Er empfiehlt eine Validierung seiner Ergebnisse anhand konkreter Inhalte.

Altersstufen und Populationen

Die Altersstufen, über die im Rahmen des Peer Tutorings berichtet wird, sind unterschiedlich. So referiert Cohens Metastudie (1982) über Studien zu den Schulstufen eins bis neun, also über Schulanfänger/innen bis hin zur Sekundarstufe 1. Rohrbecks Studie (2003) widmet sich Grundschüler/innen aus den USA und legt hier vermehrt das Augenmerk auf Schüler/innen aus Risikogruppen. Es werden hier explizit Angehörige von Minderheiten genannt und Kinder, die in städtischer Armut aufwachsen. In eine ähnliche Kerbe schlagen auch Robinson et al. (2005), die sich ebenfalls auf Minderheiten konzentrieren, nämlich auf afro-amerikanische Schüler/innen, allerdings auf Ältere aus der Sekundarstufe. Die Motivation beider Studien war es, einem unterdurchschnittlichen Abschneiden eben dieser Schülergruppen entgegen zu wirken. Fogarty und Wang wiederum beschreiben Tutoringprozesse, bei denen die Tutoren aus der Mittelstufe (sechste bis achte Schulstufe) sind und die Tutees aus unterschiedlichen Schulstufen, vom Kindergarten bis zur fünften Schulstufe. Zinn (2009) arbeitete mit Schüler/innen aus der Sekundarstufe, die Grund- und Vorschüler/innen betreuten. Einer gänzlich anderen Altersstufe widmet sich Topping (1996), nämlich Studierenden an Colleges.

Alles in allem kann zusammengefasst werden, dass Studien zu der Altersgruppe der zehn- bis 14-Jährigen zwar vorhanden, aber unterrepräsentiert sind, während solche zu Grundschulkindern und Studierenden an Colleges in einer größeren Zahl vorliegen.

2.2.4. Tutor-Tutee-Interaktionen und Altersdifferenz

Was macht nun die Art der Tutor-Tutee-Interaktion aus? Wieso sollte sie auch nur annähernd gleich gut, wenn nicht sogar besser sein, als eine Lehrer-Schüler-Interaktion? Immerhin haben Lehrer/innen eine jahrelange Ausbildung in fachlicher, pädagogischer und fachdidaktischer Hinsicht durchlaufen, während das auf die Tutoren keinesfalls zutrifft. Die Antwort ist wohl am ehesten darin zu finden, dass die Form der Interaktion zwischen Tutoren und Tutees völlig anders abläuft als die zwischen Lehrer/innen und Schüler/innen. Die Tutor-Tutee-Interaktion ist eine auf Augenhöhe, allein schon aus der Tatsache, dass es sich um Peers handelt, also um Gleichgesinnte (Fogarty & Wang, 1982). Es sind Jugendliche mit ähnlicher Sozialisation, weil sie z.B. aus derselben Schule stammen und, weil sie als Schüler/innen ähnliche Perspektiven und Sorgen haben. Die Tutor-Tutee-Interaktion basiert auf einem freundschaftlichen Umgang, im Unterschied zu der hierarchischen Lehrer-Schüler-Beziehung (Robinson, et al., 2005). Fogarty und Wang (1982) analysierten in ihrer Studie den Tutoring Prozess auf Basis verbaler Interaktionen und nutzten Ergebnisse der Rollentheorie zur Deutung. Sie geben an, dass es sich hier um ein wechselweises Geben und Nehmen zwischen den Lernpartnern handelt.

Damit diese soziale, emotionale und vor allem sprachliche Nähe, die Basis des Lernerfolges ist, nicht zerstört wird, darf der Altersabstand zwischen Tutoren und Tutees notwendiger Weise nicht zu groß sein. Robinson et al. (2005) sprechen davon, dass Tutoring-Programme, bei denen der Altersabstand zwischen zwei und vier Jahren beträgt, jedenfalls Erfolg versprechend sind. Bei größeren Altersabständen, etwa sechs oder sieben Jahre, ist Vorsicht geboten, da sie zumindest für die Tutoren wenig bringen.

Am anderen Ende der Skala steht ein Altersabstand von null Jahren, was dem Peer Tutoring (auf gleicher Altersstufe) entspricht. Diese Form des Tutorings ist im Rahmen der hier vorgestellten Studie gelegentlich auch vorgekommen und es stellt sich die Frage, was auf Grund des fehlenden Altersunterschiedes zu erwarten ist. Dazu kann man in der Literatur die Angabe finden, dass die beobachteten Effektgrößen ein wenig kleiner

sind als beim CAPT, der fehlende Altersabstand die Ergebnisse sonst unbeeinflusst lässt (Hattie, 2009).

2.2.5. Implikationen aus vorangegangenen Studien

Die oben beschriebenen Studien geben wertvolle Implikationen für die Durchführung von Tutoring Programmen und deren Beforschung betreffend einiger struktureller Parameter.

Tutoren

Zunächst sei festgehalten, dass die Tutoren vermehrtes Augenmerk verdienen, auch wenn Tutoring Programme ursprünglich zur Verbesserung der Fähigkeiten von Tutees erdacht worden sind. Aus der Literatur ist aber, wie oben geschildert, bereits bekannt, dass die Tutoren mehr als die Tutees profitieren (Fogarty & Wang, 1982; Robinson, et al., 2005; Topping, 1996). Deshalb ist die Wirkung von CAPT auf die Tutoren besonders zu beachten.

Strukturiertheit der Programme

Die Strukturierung der Programme wird von Topping (2005) und Cohen (1982) als besonders wichtig eingeschätzt. Er schreibt, *„...results are typically very good..."* (S 635), wenn das Tutoring sorgfältig strukturiert und organisiert ist und vor allem, wenn die Methode genau auf den zu behandelnden Kontext abgestimmt ist. Unter Strukturierung ist zu verstehen, ob Überlegungen unter anderem zu folgenden Punkten angestellt wurden:

- Welche Inhalte sollen bearbeitet werden?
- Mit welchen Materialien und Methoden sollen diese Inhalte bearbeitet werden?
- Wie sieht der zeitliche, räumliche und lehrplanmäßige Rahmen aus?
- Wie erfolgt die Auswahl der Tutoren? Werden Einzelne bestimmt oder wird ein klassenweites Tutoring angestrebt?
- Wie erfolgt die Zuteilung der Tutoren zu ihren Tutees? Zufällig, nach Geschlecht oder nach Fähigkeiten?

Zwar gibt es keine Hinweise, dass unstrukturierte Programme zu keinem Erfolg führen, jedoch ist ein geringerer zu erwarten, weshalb diese Punkte bei der Planung sorgfältig bedacht werden sollten (ebd.).

Praktische Durchführung

Wertvolle Folgerungen für eine Planung und praktische Durchführung von Peer Tutoring Interventionen liefert die Publikation von Robinson et al. (2005), weshalb sie hier abschließend im Überblick wiedergegeben werden.

Folgerung 1: Peer und Cross-Age Peer Tutorings sind wegen ihres Potenzials, die kognitiven (zumindest für Mathematik) wie die nicht-kognitiven Fähigkeiten zu entwickeln, ernsthaft in Betracht zu ziehen.

Folgerung 2: Obwohl Peer und Cross-Age Peer Tutoring Programme ursprünglich gedacht waren, um die Fähigkeiten der Tutees zu verbessern, soll berücksichtigt werden, dass auch die Tutoren in einem hohen Maße profitieren.

Folgerung 3: Als Tutoren müssen nicht nur leistungsfähige Schüler/innen ausgewählt werden, auch mittelmäßige oder sogar Risikoschüler/innen profitieren von der Rolle als Tutoren.

Folgerung 4: Reziprokes Tutoring, das heißt eines, wo die Rollen zwischen Tutoren und Tutees im Laufe des Tutorings getauscht werden, sollte in Betracht gezogen werden.

Folgerung 5: Die Altersdifferenz zwischen Tutoren und Tutees sollte nicht zu groß sein. Empfohlen werden zwei bis vier Jahre, damit alle Beteiligten in beiden Rollen davon profitieren.

Folgerung 6: Um die Effekte zu vergrößern, sollten gleichgeschlechtliche Tutor-Tutee-Paarungen in Betracht gezogen werden.

Folgerung 7: Es wird betont, dass Peer und Cross-Age Peer Tutoring Prozesse dazu führen können, dass Schüler/innen innerhalb ihrer Klassen neue Rollen einnehmen und eventuell vorhandene feste Muster verlassen.

Diese Implikationen wurden in Rahmen des Forschungsdesigns der vorliegenden Arbeit weitgehend umgesetzt, um das Potenzial der Methode möglichst gut auszuschöpfen und die Effekte zu maximieren.

2.3. Motivation

Neben persönlichen Fähigkeiten einer Person entscheidet die Motivation über den Lernerfolg. Um dieses Konstrukt erfassen zu können, entwickelte sich in der pädagogischen Psychologie das Gebiet der Motivationsforschung mit einer schier unüberschaubaren Vielzahl an Modellen. Auf den kleinsten gemeinsamen Nenner gebracht bezeichnen alle ein hypothetisches Konstrukt, genannt Motivation, das zielgerichtetes Handeln zur Folge hat, um einen positiv besetzten Zielzustand zu erreichen (Rheinberg, 2006). Die einzelnen Ansätze beschreiben in unterschiedlicher Ausprägung Richtung, Ausdauer und Intensität dieses Handelns (Urhahne, 2008). Ausgangspunkt ist immer ein psychologisches Bedürfnis, das durch eben diese Handlung befriedigt wird. Die klassische Psychologie spricht von Motivation, wenn persönliche Merkmale (Motive) mit passenden Situationsmerkmalen zusammentreffen (situative Motivanregung). Will man daher im Zusammenhang mit Schule und Lernen eine Person motivieren, kann man entweder die Motive einer Person, die jedoch relativ stabil sind, oder die Situationsmerkmale verändern, damit beide zueinander passen. Beschränkt man sich auf die Situationsmerkmale, heißt das konkret, dass man Lernumgebungen an die jeweilige Motivstruktur der Lerner anpasst. Da die Lernende im schulischen Zusammenhang jedoch unterschiedlich sind, wird das nie vollständig gelingen und es kann daher auch keine universellen Konzepte für motivierende Lernumgebungen geben (Rheinberg, 2006).

Im Folgenden soll nun ein kurzer Überblick über einige Motivationstheorien gegeben werden, die im Kontext von Lernen Anwendung finden. Darüber hinaus soll begründet werden, welche Theorie im Rahmen dieser Studie am besten geeignet erscheint, um die beim Cross-Age Peer Tutoring ablaufenden Prozesse erfassen zu können.

Überblick über einige Motivationstheorien

Die Leistungsmotivation beschreibt ein Verhalten, das sich an einem externen Maßstab orientiert. An ihm werden Erfolg und Misserfolg gemessen. Dieser Maßstab kann eine Schulnote sein, eine Laufzeit bei einem Rennen oder pekuniärer Gewinn. Modelliert wurde dieses Konstrukt im Risiko-Wahl-Modell von Atkinson (Atkinson, 1975). Er untersuchte zunächst die Aufgabenwahl von Proband/innen. In diesem Modell steuern

die Hoffnung auf Erfolg und die Angst vor Misserfolg das Verhalten. Das jeweilige Leistungsmotiv ist hier das persönliche Merkmal, die situative Motivanregung kommt aus den Anreizen, die eine Situation schafft.

Da es immerhin eine durchschnittliche Korrelation (Werte zwischen 0,26 und 0,31) (Urhahne, 2008) zwischen Leistungsmotivation und tatsächlichen (schulischen) Leistungen gibt, kann dieses Modell auch zu deren Prognose herangezogen werden. Atkinson ging in seinem Modell von der Annahme aus, dass der Anreizwert von Erfolg linear steigt, wenn die Erfolgswahrscheinlichkeit sinkt. Diese Linearität stellt aber ein sehr simples Modell dar.

Die Theorie der Attribution geht von Meinungen und Überzeugungen aus, die Menschen zur Erklärung bestimmter Ereignisse heranziehen. Weiner et al. (Urhahne, 2008; Weiner et al., 1971) geben hierzu vier Dimensionen an: internale und externale Gründe sowie stabile und variable Gründe.

Demnach gibt es, was die Aufrechterhaltung der Motivation einer Person anlangt, günstige und ungünstige Attributionsstile: Günstig ist es, wenn Erfolg stabilen, internalen Faktoren zugeschrieben wird, die in der Person selbst verortet sind. Misserfolg hingegen sollte variablen, externalen Faktoren, wie z.B. mangelnder Vorbereitung, zugeschrieben werden. Jedenfalls erweist sich eine individuelle Bezugsnormorientierung als sinnvoll, da attributionale Theorien zwischen Lernergebnissen und Lernprozessen eine Verbindung herstellen. Im schlechtesten Fall der Attribuierung kann es zur erlernten Hilflosigkeit kommen, die auftritt, wenn Schüler/innen Misserfolge auf stabile, internale Gründe zurückführen.

Dweck und Reppucci (1973) stellten durch Untersuchungen zu erlernter Hilflosigkeit fest, dass Schüler/innen mit gleichen Fähigkeiten unterschiedlich auf Misserfolge reagieren. Zur Erklärung führten sie das Konstrukt der Zielorientierung ein: Es gibt Lernziele und Performanzziele. Will eine Person Lernziele erreichen, versucht sie besser zu werden, will sie Performanzziele erreichen, versucht sie anderen gegenüber besser dazustehen. Es zeigt sich, dass die Lernzielorientierung gegenüber der Orientierung an Performanzzielen Vorteile für das Lernen und für die Bewältigung von Misserfolgen hat (Urhahne, 2008).

Die Theorie der Volition widmet sich der Beschreibung einer Handlung, von der Handlungsabsicht, bis zum (erfolgreichen) Durchführen und Beenden. Das Risiko-Wahl-Modell von Atkinsons besagt lediglich, dass dem stärksten Motivationsreiz gefolgt wird. Wenn dies die Motivation ist, bei Schönwetter schwimmen zu gehen, bleibt das Lernen auf der Strecke. Die Volition beschreibt nun Prozesse, die anschließend an die erste Phase der Motivierung auftreten und die durch den Willen auch Handlungsbarrieren überwinden lassen, um zum Ziel zu gelangen (Urhahne, 2008).

Die soziale Motivation beschreibt die Orientierung an Mitschüler/innen, Eltern oder Lehrer/innen zu denen versucht wird, gute Beziehungen aufzubauen oder deren Bestätigung zu erhalten. Lernleistungen von Schüler/innen korrelieren immerhin mit r = 0,4 mit den Erwartungen der Eltern (Urhahne, 2008). Die Theorie der sozialen Motivation wird mit dem Konstrukt der sozialen Eingebundenheit im Rahmen der *Self Determination Theory* nach Deci und Ryan (siehe S 34) weitgehend abgedeckt und daher hier nur der Vollständigkeit halber erwähnt.

Die Erwartungswerttheorien gehen davon aus, dass Handlungen bewertet werden und zusätzlich die Wahrscheinlichkeit, eine Handlung auch umsetzen zu können, abgeschätzt wird. Stehen mehrere Handlungsalternativen zur Verfügung, wird jene gewählt, bei der das Produkt aus Erwartung und Wert maximal ist (Urhahne, 2008). Bandura (1977) entwickelte seine *Self-Efficacy* Theorie (Theorie der Selbstwirksamkeitserwartung), indem er einerseits das behavioristische Reiz-Reaktionsschema um kognitive Faktoren zu einem sozial-kognitiven Modell erweiterte. Andererseits widersprach er ihm, wonach das Verhalten nur durch unmittelbare Konsequenzen gesteuert wird. Stattdessen postulierte er, dass das Verhalten mit dem Ergebnis, und zwar auf einem aggregierten Niveau, zusammenhängt (ebd., S 192). In weiterer Folge unterschied er zwischen Ergebniserwartung (*outcome expectations*), das ist die Einschätzung wahrscheinlicher Konsequenzen einer Handlung, und Wirksamkeitserwartung (*efficacy expectations*), das ist die Einschätzung der eigenen Fähigkeiten[5]. Eine hohe Ergebniserwartung (z.B. führt Lernen zum Erfolg bei einer Prüfung?) ist allerdings nicht ausreichend, um eine Handlung zu initiieren. Ist nämlich die Wirksamkeitserwartung zu gering, traut sich daher der

[5] *„An outcome expectancy is defined as a person's estimate that a given behavior will lead to certain outcomes. An efficacy expectation is the conviction that one can successfully execute the behavior required to produce the outcomes"* (Bandura, 1977, S. 193).

Lerner nicht zu, dass er es schaffen kann, bleibt die Handlung (in diesem Fall das Lernen) aus. Kritisch kann zu diesem Modell angemerkt werden, dass hier die Einschätzungen rein kognitiv vorgenommen werden. Die *Self-Efficacy* Theorie beschreibt somit nur *eine* Komponente der Motivation und berücksichtigt nicht die Wertedimension, die Qualität der sozialen Beziehungen und das daraus resultierende emotionale Erleben. Sie stellt daher nur eine notwendige Bedingung für das Entstehen von Lernmotivation dar, ist aber alleine keine hinreichende Erklärung. Es wird zwar die Stärke der Motivation beschrieben, aber nicht ihre Ausrichtung, w*arum* eine Handlung gesetzt wird (Krapp & Ryan, 2002). Motivation wird also als Merkmal einer Person beschrieben, das eine stärkere oder eine weniger starke Ausprägung haben kann. Woher diese Motivation kommt, ob von außen, z.B. durch starken Druck von Lehrpersonen oder Eltern, oder von innen bleibt dabei ebenso unberücksichtigt, wie tiefer liegende Gründe für ein von innen gesteuertes, motiviertes Handeln. Damit ist die *Self-Efficacy* Theorie nicht in der Lage, intrinsisch motiviertes Handeln abzubilden. Intrinsische Motivation findet ihre Ursache mehr im Handeln selbst, als in der Einschätzung der Folgen des Handelns. Daher kann eine rein kognitive Beschreibung wie die der *Self-Efficacy* Theorie nicht vollständig sein.

Trotzdem muss man der *Self-Efficacy* Theorie der Motivation zugute halten, dass sie eine hohe prognostische Valenz in vielen Bereichen besitzt, hohe praktische Bedeutung auch über das Lernen hinaus hat und über eine große Erklärungskraft verfügt. Ihre Bedeutung liegt vor allem in der Beschreibung von Handlungsgeschehen (Krapp & Ryan, 2002).

Ein Fragebogeninstrument, basierend auf dem theoretischen Konstrukt der *Self-Efficacy* Theorie stellt der in Kapitel 5.3.3 beschriebene Fragebogen zur allgemeinen Selbstwirksamkeitserwartung (Schwarzer & Jerusalem, 1999b) dar.

Die Selbstbestimmungstheorie der Motivation (SDT)

Die bereits oben kurz erwähnte *Self Determination Theory* (SDT), die Theorie der Selbstbestimmung, nach den Ideen von Deci und Ryan, stellt in gewissem Sinne eine Erweiterung der *Self-Efficacy* Theorie dar, indem die Ursachen und somit die Qualität der Motivation erfasst werden. Dabei spielt die Beschreibung intrinsischer Motivation eine zentrale Rolle, wobei die SDT es sich zum Ziel setzt, die sozialen und umfeldbedingten Faktoren zu identifizieren, die intrinsische Motivation entweder fördern, oder unterminieren (Deci & Ryan, 2008; Ryan & Deci, 2000).

Die SDT legt nahe, dass unterschiedliche Ursachen für ein motiviertes Verhalten auch zu Unterschieden in der längerfristigen Aufrechterhaltung einer Handlung und in der Qualität des emotionalen Erlebens führen. Diese Unterschiede wiederum beeinflussen die Handlungsergebnisse. Im Fall der Lernmotivation bedeutet dies, dass durch äußere Anreize motiviertes Lernen nur so lange aufrecht erhalten wird, so lange der äußere Anreiz vorhanden ist und als genügend stark empfunden wird. Untersuchungen konnten zeigen, dass eine Motivierung der Lernenden von außen zu oberflächlicheren Strategien in der Aufgabenbearbeitung führt, die Kreativität nicht fördert und allgemein der einfachste Weg gesucht wird, um ein Problem zu lösen (Krapp & Ryan, 2002). Zusätzlich verbinden die Lernenden mit durch äußere Anreize allein motivierten Aufgaben oft negative Gefühle, wie Stress oder Angst und/oder Widerwillen. Das gibt im Rahmen der SDT Anlass, die Ursachen der Motivation differenziert zu betrachten, um damit auch ihre emotionale Qualität beschreiben zu können. Deci und Ryan wählen dazu die Beschreibung über die Anreizbedingungen und unterscheiden, ob ein Verhalten aus der Person selbst (intrinsisch) motiviert ist oder von außen (extrinsisch) motiviert ist.

Zunächst sei festgehalten, dass es neben extrinsischer und intrinsischer Motivation auch amotiviertes Verhalten gibt. Es ist ein Regulationsstil, der durch das Fehlen von Autonomie gekennzeichnet ist. Amotivierte Personen sehen nicht, dass ihr Handeln mit Resultaten kausal verknüpft ist oder dass notwendiges Handeln im Rahmen ihrer Möglichkeiten oder in ihrer Umgebungen durchführbar ist. Eine amotivierte Person fühlt sich weder autonom, noch kompetent, es fehlt ihr vollständig an Verinnerlichung von Werten und Normen (Ryan, 1995). Daher wird Amotivation in Abbildung 2.1 auch als getrennte Kategorie geführt.

Die SDT geht, wie alle Motivationstheorien, davon aus, dass motiviertes Handeln auftritt, wenn psychologische Bedürfnisse befriedigt werden wollen. Im konkreten Fall werden im Rahmen dieser Theorie drei *basic needs* angegeben: Autonomieerleben, Kompetenzerleben und soziale Eingebundenheit.

Die Geschichte der drei *basic needs* geht ursprünglich auf Forschung an Tieren in den 1950er-Jahren zurück, als man spontane, tierische Verhaltensmuster entdeckte, die scheinbar ohne äußere Einwirkung auftraten. Die Erforschung der Umstände, unter denen dieses Verhalten auftrat, führte zur Identifizierung der drei *basic needs* (Ryan,

1995), die auch für menschliches Verhalten ein äußerste leistungsfähiges Konzept darstellen. Erstaunlicherweise stellte sich heraus, dass diese drei *basic needs* kulturunabhängig sind (Deci & Ryan, 2008).

Das Autonomieerleben ist vielleicht am engsten mit der intrinsischen Motivation verbunden: Sie tritt nur dann auf, wenn sich ein Mensch in seinem Handeln als autonom erlebt, als nicht von außen gesteuert (Ryan & Deci, 2000). Autonomieerleben ist umso größer, je mehr Entscheidungsmöglichkeiten es gibt, je geringer der Druck von außen ist (Ryan, 1995) und der Mensch das Gefühl hat, eine Stimme, die gehört wird, und eine Wahlmöglichkeit zu haben (*voice and choice*). Das ist durch zahlreiche empirische Untersuchungen im Rahmen der *Cognitive Evaluation Theory* (Deci & Ryan, 1985; Ryan, 1982), die als Subtheorie der SDT aufgefasst werden kann, bestätigt worden (Krapp & Ryan, 2002). Dieses Autonomieerleben ist jedoch nicht mit einer Beliebigkeit zu verwechseln, dass jemand tun und lassen kann, was er will. Es handelt sich hier um die innere Übereinstimmung zwischen dem, was ein Mensch für wichtig hält (und daher tun möchte) und dem, was die aktuelle Situation erfordert (Krapp & Ryan, 2002).

Das Kompetenzerleben bezeichnet das Bedürfnis eines Menschen sich in einer Situation selbst als kompetent wahrnehmen zu können. Hier trifft das Konstrukt der SDT jenes der *Self-Efficacy Theory*: Kompetenzerleben in der einen Theorie wird in der anderen Theorie als Wirksamkeit bezeichnet. Kompetenz wird erlebt, wenn Lernende im Rahmen der ihnen gestellten Aufgaben die optimale Herausforderung erfahren. Eine Aufgabe darf nicht zu leicht sein, aber auch nicht so schwer, dass sie im Rahmen der eigenen Wirksamkeit als unlösbar empfunden wird. Das fügt sich auch in das Bild, das konstruktivistische Lerntheorien zeichnen, wonach die Passung der Aufgabenstellungen eine entscheidende Rolle für eine konzeptuelle Weiterentwicklung der Lernenden spielt. Eine weitere wesentliche Rolle spielt positives Feedback (Ryan, 1995).

Das dritte psychologische Grundbedürfnis ist die soziale Eingebundenheit. Darunter ist eine mehr oder minder starke, positive Verbundenheit mit „dem Anderen“ zu verstehen. „Der Andere“ ist oft jemand aus der Peergroup, jemand, der jedenfalls persönliche Bedeutung besitzt. Eingebundenheit resultiert auch aus einem als sicher erlebten sozialen Hintergrund (Ryan, 1995). Wird diese Verbundenheit wahrgenommen, ist sie ein zentraler Bestandteil der persönlichen Motive, die zu motiviertem Verhalten führen.

Die Eingebundenheit operationalisiert die Auffassung konstruktivistischer Lerntheorien, wonach Lernen immer vor einem sozialen Hintergrund und in einem kulturellen Kontext stattfindet.

Die soziale Eingebundenheit als Prädiktor für intrinsische Motivation besitzt nicht jene Bedeutung, die Autonomie und Kompetenzerleben haben. Es gibt durchaus Tätigkeiten, die jemand alleine durchführen kann und dabei trotzdem intrinsische motiviert ist, also die Handlung um ihrer selbst willen durchführt (Deci & Ryan, 2000). Jedoch ist die soziale Eingebundenheit etwas, das Internalisationsprozesse und somit eine Entwicklung von extrinsischer zu intrinsischer Motivation unterstützt (Ryan & Deci, 2000).

Das Autonomieerleben stellt im Rahmen der SDT den entscheidenden Parameter dar, ob intrinsische oder extrinsische Motivation vorliegt. Diese beiden Ausprägungen der Motivation sind als zueinander komplementär zu betrachten. In Bereichen, wo die intrinsische Motivation nicht ausreicht, um ein Verhalten aufrecht zu erhalten, z.B. wenn es innerhalb von Spracherwerb um Vokabellernen geht, können extrinsische Anreize von Zeit zu Zeit von Hilfe sein, auch für an sich intrinsisch motivierte Personen. Die Bedingung ist allerdings, dass dieses extrinsische Verhalten als zumindest teilweise selbstbestimmt wahrgenommen wird. Die extrinsische Motivation hat dann keine negativen Auswirkungen auf die intrinsische, sondern dient zu ihrer Aufrechterhaltung (Deci & Ryan, 1993). Diese Erkenntnis veranlasste Deci und Ryan zu einer weiteren Differenzierung innerhalb der extrinsischen Motivation.

Intrinsische und extrinsische Motivation in der SDT

„Intrinsic motivation is defined as the doing of an activity for its inherent satisfaction rather than for some separable consequences“ (Ryan & Deci, 2000, S. 56). Intrinsische Motivation beruht daher darauf, dass eine Handlung an sich bereits als erfüllend empfunden wird, weil die Tätigkeit an sich Freude bereitet oder weil es ein intrinsisches *Interesse* an einer *Sache* gibt. Es bedarf keiner weiteren Handlungsanreize von außen, wie in Aussicht gestellter Belohnungen oder angedrohter negativer Konsequenzen. Intrinsisch motivierte Handlungen sind ein Ausdruck des Selbst einer Person, mit einem

internal perceived locus of causality[6], einer ausschließlich im Selbst wahrgenommenen Ort der Entscheidungsfindung (Ryan, 1995). Aus diesem Grund ist in Abbildung 2.1 die intrinsische Motivation auch als getrennte Kategorie dargestellt.

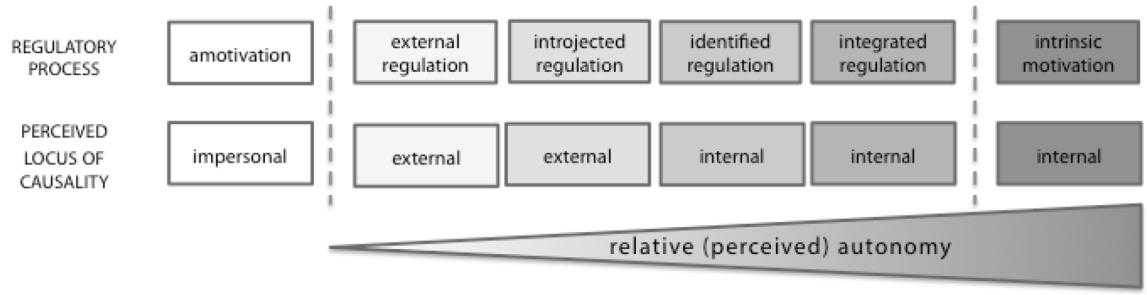


Abbildung 2.1: Überblick über die Arten der (A-) Motivation, den Ort der Ursächlichkeit und die relative, wahrgenommene Autonomie (nach: Ryan, 1995, S. 406, fig. 1).

Das Konstrukt der intrinsischen Motivation stellt insofern eine wesentliche Erweiterung der Leistungsmotivation nach Atkinson dar, als hier ganz andere Motive den Ausschlag geben. In der Auffassung der Leistungsmotivation sind es alleine von außen festgelegte Maßstäbe, die motiviertes Verhalten erzeugen. Es existiert hier keine Selbstbestimmung des Individuums bei ihrer Festlegung. In der Auffassung der intrinsischen Motivation fehlen diese äußeren Handlungsanreize völlig, ja sie sind sogar aufgrund der fehlenden Autonomie kontraproduktiv.

Es gibt einige empirische Untersuchungen, die nachweisen, dass intrinsische Motivation ein Bedingungsfaktor ist, der zu qualitativ anspruchsvollen Formen des Lernens führt (Krapp & Ryan, 2002). Die *Cognitive Evaluation Theory* (Deci & Ryan, 1985; Ryan, 1982), interpretiert und fasst Befunde zusammen, wonach Kontrollmaßnahmen und ungeeignete Lernumgebungen, die das Gefühl von Inkompetenz und daher geringer Wirksamkeit vermitteln, intrinsische Motivation konterkarieren, weil der wahrgenommene Ort der Entscheidung von innen nach außen verlagert wird. Hingegen sind Autonomieunterstützung und ein Feedback, das auf eine Unterstützung der Wirksamkeitserwartung abzielt, der intrinsischen Motivation förderlich. Die Rolle von Belohnungen und Lob ist umstritten, da es beides bewirken kann: Werden sie als

[6] Der Begriff des *internal perceived locus of causality* wurde von deCharms in seinem Buch „Personal causation: The internal affective determinants of behavior“ (1968, Reprint 2013, Routledge, New York) geprägt und von Deci und Ryan übernommen (1985).

Kontrolle empfunden, untergraben sie die intrinsische Motivation, werden sie als Feedback empfunden, wirken sie unterstützend.

„Extrinsic motivation is a construct that pertains whenever an activity is done in order to attain some seperable outcome“ (Ryan & Deci, 2000, S. 60). Somit steht die extrinsische Motivation im Kontrast zur intrinsischen, weil bei letzterer Dinge um ihrer selbst Willen gemacht werden, ohne dass es einen instrumentellen Wert gibt.

Innerhalb der extrinsischen Motivation unterscheiden Deci und Ryan vier Regulationsmechanismen: die externale (*external*), die introjizierte (*introjected*), die identifizierte (*identified*) und die integrierte (*integrated*) Regulation. Diese vier Formen unterscheiden sich im Grad der (wahrgenommenen) Autonomie einer Person, von den Endpunkten „äußere Kontrolle“ bis „Selbstbestimmung“.

Im Rahmen der SDT werden zwei Prozesse beschrieben, die extrinsisch motivierte Verhaltensweisen in intrinsische überführen können: Es ist dies die Internalisation und die Integration von Werten und Normen. Unter Internalisation versteht man den Prozess, bei dem externale Werte und Normen in internale Regulationsprozesse einer Person aufgenommen werden. Integration ist insofern die Fortsetzung davon, als dass die durch den internalen Regulationsstil aufgenommenen Werte und Normen ganz ins Selbst eingebunden werden. (Ryan & Deci, 2000).

Externale Regulation liegt vor, wenn Handlungen ausgeführt werden, um eine Belohnung zu bekommen oder einer Strafe zu entgehen, also von äußeren Anreizen gesteuert sind. Als Beispiel sei ein Schüler genannt, der eine Hausübung ausschließlich macht, um keine schlechte Note zu bekommen. Es liegt hier weder Freiwilligkeit noch Autonomie vor.

Introjizierte Regulation beschreibt ein Verhalten, das einem inneren Druck folgt, für die Selbstachtung relevant ist oder bei Unterlassung ein schlechtes Gewissen erzeugen würde. Zwar sind hier keine äußeren Handlungsanstöße nötig, aber es handelt sich dennoch um eine erzwungene Handlung, mit dem Unterschied zu oben, dass Regulator und Regulierter zwar zwei Personen sind, ihre Werte aber in einer Person wohnen. Für diese Person wird der *perceived locus of causality* extern empfunden (Deci & Ryan,

1993). Ein Beispiel dafür wäre ein Schüler, der die Hausübung macht, weil er sich sonst vor Eltern oder Lehrer/innen schämt.

Von identifizierter Regulation spricht die SDT, wenn auch das Selbst ein Verhalten als persönlich wichtig und wertvoll erkennt. Um beim Beispiel der Schülerin zu bleiben, macht sie die Hausübung nicht nur, weil sie keine schlechte Note bekommen möchte, sondern weil sie persönlich ihre Wichtigkeit z.B. für den persönlichen Lernfortschritt erkennt. Der Wert (hier: Hausübung machen) wurde erkannt und in das Selbstkonzept integriert (Deci & Ryan, 1993).

Die integrierte Regulation ist der intrinsischen durch den hohen Grad an Selbstbestimmung am ähnlichsten, sie ist die eigenständigste Form der extrinsischen Motivation. Es findet nicht nur eine Identifizierung mit Zielen und Normen statt, sie werden sogar kohärent in das Selbst integriert.

Integrierte Regulation und intrinsische Motivation bilden in der SDT gemeinsam die Basis selbstbestimmten Handelns. Das ist eine der zentralen Unterscheidungen, die die SDT macht, nämlich die zwischen kontrollierten und autonomen Formen der Motivation (Deci & Ryan, 2008).

Die SDT setzt sich zum Ziel nicht nur die Qualitäten und Ursachen der Motivation zu beschreiben, sondern auch Möglichkeiten aufzuzeigen, wie extrinsische Motivation in intrinsische transformiert werden kann. Dazu ist es erforderlich, dass Verhaltensweisen und Lernumgebungen identifiziert werden, die die intrinsische Motivation unterstützen und nicht unterbinden. Eine der zentralen Aussagen der SDT ist es, dass neben dem Kompetenzerleben vor allem die Autonomieunterstützung entscheidend ist.

Cross-Age Peer Tutoring und Motivationstheorien

Folglich wird zu untersuchen sein, ob Cross-Age Peer Tutoring als Unterrichtsmethode, mit dem so designten Mentoring (vgl. Kapitel 4.4), das durch die Wahlmöglichkeit bei den Aufgaben und durch den Expertenstatus der Tutoren für das Lernen der Tutees die Autonomie der Schüler/innen unterstützt, die Entwicklung autonomer, selbstregulierter Motivationsformen fördert.

Nach Durchsicht der hier vorgestellten Motivationstheorien scheint es, dass die SDT jene Theorie ist, die am besten mit den Ideen konstruktivistischer Lerntheorien in Einklang zu

bringen ist. Verlockend sind allerdings auch die Befunde zur *Self-Efficacy Theory*, wonach sie sich als guter Prädiktor für Lernleistungen nutzen lässt. Letztlich sind wir im Rahmen einer Studie über die Wirksamkeit einer Unterrichtsmethode (CAPT) hoch interessiert an diesen. Dennoch vermag die SDT zielgerichteter die Lernprozesse beim CAPT zu beschreiben. Darüber hinaus stimmen Teile ihres Motivationskonstrukts weitgehend mit dem Konstrukt der *Self-Efficacy Theory* überein. So werden die *outcome expectations* im Rahmen der Autonomie einer Person abgeschätzt. Die *efficacy expectations* bilden einen Teil des *perceived competence* Konstrukts. Die SDT stellt sich somit als die umfassendere Theorie dar.

Hinsichtlich eines möglichen *Conceptual Change*, den CAPT möglicher Weise initiieren kann, ist es vor dem Hintergrund einer Erweiterung des klassischen Ansatzes zu einem multi-perspektivischen Modell (Duit & Treagust, 2003) sinnvoll, die Motivation zu erfassen. Hinsichtlich einer Implementierung von CAPT als Unterrichtsmethode in den Regelunterricht ist ebenfalls sinnvoll entscheiden zu können, ob es sich hier um eine Methode handelt, die durch ihren Aufbau alleine, vielleicht sogar weitgehend unabhängig von den gewählten Themen, motivierend auf Schüler/innen wirkt, da Physikunterricht oft als unbeliebt, ja sogar abschreckend empfunden wird (Muckenfuß, 1995). Die spezielle Wirkung von CAPT auf die Motivation kann darin bestehen, dass die psychologischen Grundbedürfnisse (*basic psychological needs*) im Sinne der SDT (Deci & Ryan, 2008) angesprochen und zufrieden gestellt werden, was motivierend wirkt. Da die Erfassung der Motivation eine entscheidende Grundlage für die Beurteilung von CAPT darstellt, ist es nötig auch diesen nicht-kognitiven Parameter beschreiben zu können.

3. Forschungsfragen

Im Kapitel 2 wurde anhand der diskutierten Literatur ein ausführlicher Überblick über mögliche Arten der Durchführung und Evidenzen zur Wirkung von CAPT als Unterrichtsmethode hinsichtlich unterschiedlichster Themen und bezüglich verschiedener Altersstufen sowie im Zusammenhang mit motivationalen Aspekten gegeben.

Aus der Sicht der Physikdidaktik im Speziellen und der Altersstufe, über die hier berichtet wird, ergeben sich in diesem Bild jedoch einige Lücken, die nun aufgezeigt werden und zu konkreten Forschungsfragen führen.

3.1. Forschungslücken

Aus der oben diskutierten Literatur ergibt sich das Bild, dass die Lernwirksamkeit von CAPT hinsichtlich einiger spezieller Themen durch Studien bereits gut untersucht ist. Dieses Spektrum an Themen betrifft in überraschend vielen Fällen Mathematik, auch im Sinne klassischer Nachhilfe, sowie Leseförderung (P. A. Cohen, et al., 1982) und die Verbesserung der Fähigkeiten im Umgang mit Computern (Fogarty & Wang, 1982). Die in Deutschland entwickelte Variante des Peer Tutorings, das Lernen durch Lehren (Martin, 1998), zielt inhaltlich hingegen auf das Erlernen und Praktizieren von Fremdsprachen ab.

Im Kontext des naturwissenschaftlichen Unterrichts, speziell des Physikunterrichts, ist die Studie von Howe et al. (1995) zu nennen, die die Einflüsse von Peer Tutoring auf das Verständnis von Heiz- und Kühlprozessen untersucht hat. Eine weitere Untersuchung aus der jüngeren Zeit ist eine Pilotstudie von Zinn (2008, 2009) zum Einsatz der Methode Lernen durch Lehren im Rahmen des naturwissenschaftlichen Unterrichts. Dabei wurden vor allem die Entwicklung des situationalen Interesses und des Interesses von Mädchen anhand unterschiedlicher Lerninhalte aus dem naturwissenschaftlichen Kontext untersucht.

Zurzeit (2014) läuft an der Universität Osnabrück in der Arbeitsgruppe von R. Berger eine Studie über Cross-Age Tutoring in Physik (M. Müller, Berger, & Hänze, 2013). Hierbei soll

zunächst die Motivation und der Wissenserwerb der Tutoren in der Vorbereitungsphase erfasst werden, um auf Basis dieser Ergebnisse ein Tutorentraining zu entwickeln, das danach evaluiert werden soll. Es ist geplant, diese Unterrichtsmethode weiterzuentwickeln, um den physikalischen Unterricht in Hauptschulen zu fördern. Ergebnisse dazu liegen aber derzeit noch nicht vor.

Neben den hier vorgestellten Studien sind keine weiteren Untersuchungen zur Lernwirksamkeit von Cross-Age Peer Tutoring im naturwissenschaftlichen Kontext bekannt.

Die Auseinandersetzung, speziell mit dem naturwissenschaftlichen Kontext, erscheint deshalb bedeutsam, da für den naturwissenschaftlichen Unterricht der Vorgang des Konzeptwechsels (*Conceptual Change*) zentral ist. Damit ist nicht gemeint, dass es in anderen Disziplinen zu keinem Konzeptwechsel beim Lernen kommt oder kommen soll, jedoch steht dieser in der Physikdidaktik seit geraumer Zeit im Zentrum konstruktivistischer Lerntheorien, die die Basis für mögliche Unterrichtskonzepte bieten. Dieser Konzeptwechsel soll die Vorstellungen, die Schüler/innen bereits in den Unterricht mitbringen, auf grundlegende Weise restrukturieren, um zu wissenschaftlich anschlussfähigen Konzepten zu führen (Duit, 1999). Dabei sind gewisse Schwierigkeiten zu erwarten, da die vorunterrichtlichen Vorstellungen, die zumeist aus dem Alltag kommen und tief verwurzel sind, nicht nur sehr stabil und manchmal lernhinderlich sind, sondern oft auch zu wissenschaftlich anschlussfähigen Konzepten in deutlichem Widerspruch stehen (Duit, et al., 2008). Vor diesem Hintergrund scheint es für eine Unterrichtsmethode wie CAPT entscheidend zu sein, ob und wie viel sie zu diesem Begriffswechsel beitragen kann.

Selbstbestimmten Lernformen wird bescheinigt, dass sie konstruktivistisch orientiertes Lernen in einer geeigneteren Form unterstützen als eine lehrerzentrierte Unterrichtspraxis (Duit, et al., 2008). Zu der Wirksamkeit kooperativer Lernformen, zu denen Peer Tutoring auch gezählt werden kann, existieren ebenfalls in einigen Kontexten Untersuchungen (Treagust, 2007). Versteht man Cross-Age Peer Tutoring als eine selbstbestimmte und kooperative Lernform, kann es den obigen Forschungsergebnissen zufolge möglich sein, dass diese Kombination von Ansätzen ebenfalls erfolgreich ist. Dennoch ist bei der Kombination beider Ansätze Vorsicht

geboten, da sich positive Effekte nicht notwendiger Weise verstärken müssen. Außerdem könnte es passieren, dass die Lernenden die Fehlkonzepte der lehrenden Schüler/innen übernehmen und diese dann, da sie von Peers stammen, umso tiefer verinnerlichen.

Eine weitere Frage von Interesse ist, ob sich CAPT für Schüler/innen der Altersstufe der 10 bis 14-Jährigen, also der Sekundarstufe 1, gut eignet. Das erscheint insofern bedeutend, als dass OECD-Studien, wie PISA 2006 und 2009 (OECD, 2010; Schreiner, 2009), offenlegen, dass österreichische Jugendliche in dieser Altersstufe signifikant unter dem Durchschnitt abschneiden. Auch verstärkt sich in dieser Altersstufe die Differenz zwischen leistungsstarken und leistungsschwachen Schüler/innen, die in der Volksschule noch nicht so ausgeprägt ist, weiter. Das zu Beginn dieser Altersstufe schon geringere Interesse der Mädchen am Physikunterricht geht im Laufe der Sekundarstufe 1 weiter zurück (Häußler, et al., 1998). Vor diesem Hintergrund scheint es umso interessanter, ob CAPT hier als Unterrichtsmethode in der Lage ist, diesen Trends entgegenzusteuern, beziehungsweise ob CAPT überhaupt für diese Altersstufe geeignet ist. In der Literatur finden sich zwar einige Studien über Peer Tutoring für die jüngeren Schüler/innen der Primarstufe, z.B. (Rohrbeck, et al., 2003) und auch für Ältere, Studierende an Colleges, (z.B. Topping, 1996), aber kaum für die so interessant erscheinende Sekundarstufe 1.

Der Ansatz Lernen durch Lehren (Martin, 1998), der auch in der Sekundarstufe 1 erprobt wurde, ist mit dem Fremdsprachenlernen thematisch in einem fundamental anderen Kontext angesiedelt. Zudem liegen keine empirischen Untersuchungen zu seiner Wirksamkeit vor. Die oben zitierte Studie von Zinn (2008) ist eine der wenigen Untersuchungen zum Lernen durch Lehren, die sowohl im naturwissenschaftlichen Kontext durchgeführt wurde, als auch empirische Befunde liefert.

Es besteht somit ein Forschungsbedarf zum Thema CAPT. In der hier berichteten Studie wurde daher auf die folgenden Forschungsfragen fokussiert.

3.2. Forschungsfragen

- Forschungsfrage 1: Weisen Schüler/innen der Sekundarstufe 1 nach der CAPT-Intervention bessere kognitive Testergebnisse auf und wenn, wie stark ist der gemessene Effekt?

  Zu Beginn der Untersuchung soll erhoben werden, wie sich CAPT auf die Entwicklung des Wissens im gesamten Sample auswirkt, noch bevor detaillierte Analysen angestellt werden. Diese unspezifische Analyse soll vor allem darüber Aufschluss geben, ob CAPT als Unterrichtsmethode für ganze Klassen geeignet erscheint, ohne zuvor Schüler/innen im Detail diagnostizieren zu müssen. Falls diese Frage positiv beantwortet werden kann und die Methode darüber hinaus noch zu ausreichenden Effektstärken führt, ist eine Implementierung von CAPT in den Regelunterricht sinnvoll. Somit kann eine praktisch bedeutsame Unterrichtsmethode, die eine konzeptuelle Entwicklung der Schüler/innen unterstützt, identifiziert und empfohlen werden.

- Forschungsfrage 2: Weisen auch die Tutoren nach der CAPT-Intervention bessere kognitive Testergebnisse auf?

  Vor den Hintergrund eines Wandels in der Begrifflichkeit von Peer Tutoring und einer damit verbundenen Verlagerung im Forschungsfokus weg von den Tutees hin zu den Tutoren, ist es auch im physikalischen Kontext interessant, ob Ergebnisse aus vorangegangenen Studien bestätigt werden können, wonach *überwiegend* die Tutoren profitieren.

- Forschungsfrage 3: Welche Unterschiede in den Posttest-Ergebnissen ergeben sich dadurch, dass die Schüler/innen innerhalb des Tutoring-Prozesses unterschiedliche Rollen inne hatten?

  Im Rahmen dieser Untersuchungen soll analysiert werden, ob sich ein Zusammenhang zwischen den unterschiedlichen Rollen, die Schüler/innen im Tutoringprozess innehaben können, und dem Posttestergebnis herstellen lässt. Diese unterschiedlichen Rollen können die Rolle als Tutor sein, die Rolle als Tutee oder aber die Doppelrolle, bei der die Proband/innen einmal als Tutee und in einem zweiten CAPT-Durchgang als Tutor tätig sind.

- Forschungsfrage 4: Welche relevanten Prädiktoren können für die Modellierung der Posttest-Ergebnisse im Rahmen einer multiplen linearen Regressionsanalyse identifiziert werden?

  Um die Ergebnisse der Forschungsfragen 1 bis 3 zusammenzufassen und die restliche vorhandene Datenbasis mit einzuschließen, wäre es wünschenswert, Prädiktoren und ihren quantitativen Einfluss identifizieren zu können, die das Abschneiden der Schüler/innen im Posttest a priori schätzen lassen. Parallel dazu ist es auch von Interesse, ob es Faktoren gibt, die keinen Einfluss auf die Posttest-Ergebnisse haben.

  Kann so ein Modell modelliert werden, ist es neben der theoretischen Eleganz auch von praktischer Bedeutung, vor allem hinsichtlich einer Implementierung von CAPT in die Schulpraxis. So ein Modell könnte Hinweise für Antworten auf die Frage nach der Wirksamkeit dieser Unterrichtsmethode für Schüler/innen mit Migrationshintergrund liefern, ebenso ob sich geschlechtsspezifische Unterschiede identifizieren lassen.

- Forschungsfrage 5: Wie nachhaltig ist die kognitive Verbesserung, die durch die CAPT-Intervention erzielt wurde?

  Betrachtet man die Lernwirksamkeit von traditionellem Unterricht (Hattie, 2009; Muckenfuß, 1995) so ist sie oft erschreckend gering. Falls überhaupt eine Lernwirksamkeit durch CAPT identifiziert werden kann, so ist es im Sinne einer Evaluationsstudie von Interesse beurteilen zu können, ob dieses Wissen Persistenz zeigt, weil es durch die intensivere kognitive Auseinandersetzung tiefer im Gedächtnis der Schüler/innen gespeichert ist.

- Forschungsfrage 6: Welche Einflüsse auf die Posttest-Ergebnisse ergeben sich aufgrund der unterschiedlichen Klassenzugehörigkeiten?

  Da es sich in dieser Studie um Feldforschung handelt, sind naturgemäß die Voraussetzungen, die die einzelnen Klassen mit sich bringen, und verschiedene Punkte, das Treatment betreffend, unterschiedlich. Auch wären für klassische Praetest-Posttest-Follow-up – Testanalysen homogenere Samples wünschenswert, was zum Beispiel die Altersstufe, den Altersabstand zwischen den Tutoren und Tutees, oder die Zeitspanne zwischen Mentoring und Tutoring betrifft. Da das aber im Rahmen der Untersuchungen nicht präziser zu

kontrollieren war, sollen Klassen einzeln analysiert werden und deren möglicherweise unterschiedliches Abschneiden analysiert werden.

- Forschungsfrage 7: Welche Zusammenhänge ergeben sich zwischen nicht-kognitiven Parametern und Testergebnissen?

  Im Rahmen dieser Forschungsfrage soll geklärt werden, ob einerseits CAPT die Motivation der Schüler/innen erhöht und ob andererseits Schüler/innen mit einer hohen Motivation oder einem guten Selbstkonzept größere Lernerfolge durch CAPT zeigen. Daneben sollen auch Analysen speziell zur Motivation von Mädchen im Rahmen von CAPT durchgeführt werden.

4. Forschungsdesign

4.1. Methodischer Zugang

Die vorliegende Arbeit entstand im Zuge eines Sparkling Science Projektes[7] und bildet einen Teil der dort geleisteten Forschungsarbeit ab. Diese Art von Projekt zielt darauf ab Forschungs-Bildungs-Kooperationen zu initiieren, bzw. zu fördern. Wunsch der Fördergeber ist es, durch das Studiendesign eine Zusammenarbeit von Schüler/innen und Wissenschaftler/innen auf Augenhöhe zu ermöglichen, um die jeweilige Arbeit wechselseitig zu befruchten.

Im Rahmen des Gesamtprojektes wurden nun die einzelnen Forschungsvorhaben so aufgeteilt, dass in der vorliegenden Arbeit der Forschungsfokus auf zwei Schwerpunkten liegt: Einerseits betrifft er die der Gestaltung der Lernumgebungen für Mentoring (das ist die Vorbereitung für Tutoren) und Tutoring (das bezeichnet die eigentliche Intervention). Andererseits sollten die Lernergebnisse, die durch das Peer-Tutoring erzielt wurden, dokumentiert werden. Dabei steht der quantitative Ansatz im Sinne einer quantitativ empirischen Fachdidaktikforschung im Vordergrund. Parallel dazu wurden im Sinne eines *mixed methods approach* (z.B. Creswell & Garrett, 2008) qualitative Studien durchgeführt. Diese finden aber in der vorliegenden Arbeit keinen Niederschlag.

Methodisch orientiert sich die Studie an den wissenschaftlichen Methoden und Standards der quantitativen Sozialforschung. Die speziellen Fragestellungen legen eine Betrachtung der Prozesse in der Tradition der Evaluationsforschung nahe, die sich zwar derselben Methoden wie die Grundlagenforschung bedient, aber einen anderen Fokus verfolgt. So sind hier die Ziele von vornherein klar, ein rein exploratives Vorgehen zur Hypothesengenerierung ist nicht nötig (Bortz & Döring, 2003). Denn als Grundlage für die Evaluation dient der theoretische Rahmen der Fachdidaktik Physik, im speziellen Fall die konstruktivistischen Lerntheorien. Auf Basis derer wurden Unterrichtsmaterialien entwickelt und die Intervention designt. Das wird in der Sprechweise der

[7] Projektnummer SPA/03 – 012/Peer Tutoring, gefördert vom ehemaligen Bundesministerium für Unterricht, Kunst und Kultur (derzeit: BM für Bildung und Frauen).

Evaluationsforschung als Maßnahme bezeichnet. Präzise gesprochen muss man die Entwicklung der Maßnahme der Interventionsforschung zuordnen. Es ist aber im Rahmen größerer Projekte sinnvoll, wenn diese beiden Teile, Interventionsdesign und Evaluation parallel ablaufen (Bortz & Döring, 2003). Die Maßnahme wurde im Sinne einer summativen, daher hypothesenüberprüfenden Evaluation evaluiert. Dank des bereits existierenden theoretischen Rahmens war es möglich, gerichtete Hypothesen zu formulieren und zu testen.

Aufgabe der Evaluationsforschung ist es, Wirkungen (den Prozess) und Wirksamkeiten (die Folgen) einer Maßnahme zu überprüfen und zu dokumentieren (Bortz & Döring, 2003; Gollwitzer & Jäger, 2009). Die beforschte Maßnahme ist hier CAPT. Das Evaluationskriterium ist dabei die Qualität des Interventionskonzeptes, um es anhand seiner Wirksamkeit auf die Lernergebnisse der beteiligten Schüler/innen zu beurteilen.

Innerhalb des Vier-Ebenen-Modells nach Kirkpatrick (2006) wurden neben der Ebene der Akzeptanz der Intervention vor allem die Ebene des Wissenserwerbs der Schüler/innen beforscht. Im Rahmen dieser Wirksamkeitsevaluation wurde beurteilt, ob durch die Intervention in der Stichprobe tatsächlich Veränderungen eingetreten sind, ob diese anhalten (Persistenz zeigen) und ob sie für die Population generalisierbar sind. Die Operationalisierung der Evaluation erfolgte durch Prae-Post und Follow-up Messungen.

Die Evaluation selbst erfolgte im Feld, daher in authentischen Situationen des Lernens. Das bringt es mit sich, dass strenge Laborbedingungen nicht eingehalten werden konnten, ja es auch gar nicht im Sinne der Experimentatoren war, solche Bedingungen zu erzeugen, da eben auch die Praxistauglichkeit von CAPT beurteilt werden sollte. Daher wurde im gesamten Projekt mit einer Vielzahl an Schülergruppen aus unterschiedlichen Schulformen und breit gestreutem Alter, ja sogar mit Vorschulgruppen aus dem Kindergarten, gearbeitet.

Mit der Festlegung der Forschungsfragen und damit mit der Fokussierung dieser Arbeit auf eine mehrheitlich quantitative Methodik, schied jedoch ein Teil der beteiligten Kinder und Jugendlichen für diese Beschreibung aus. Für die Allerkleinsten, die vor dem Schuleintritt stehen, und die kaum lesen und schreiben können, sind fragebogenbasierte Untersuchungen aufgrund des fehlenden Vermögens zum sinnerfassenden Lesen ungeeignet. Darüber hinaus sind die eingesetzten Fragebögen und Wissenstests nach

Aussage der Verfasser/innen oft erst ab einem Alter von etwa zehn Jahren geeignet (z.B. F. H. Müller, Hanfstingl, & Andreitz, 2007). Für den selbst erstellten Fragebogen zur Motivation (vgl. Kapitel 6) wurden daher auch keine Pilotierungen mit Grundschulkindern durchgeführt, um eine etwaige Eignung abzuklären.

Dies und die Auswahl an Themen für die Intervention begründet die Beschränkung der Forschungsfragen und der Auswertungen auf mögliche Effekte innerhalb der Sekundarstufe.

Wie aus der Formulierung der Forschungsfragen (Kapitel 3.2) hervorgeht, ist die Entwicklung des Wissens der Schüler/innen ein zentraler Punkt. Deshalb stehen auch die Wissenstests eindeutig im Vordergrund. Um die zeitliche Entwicklung des Wissens zu erfassen, wurde zu drei Testzeitpunkten getestet: vor der Intervention, unmittelbar nach der Intervention und einige Zeit später. Es liegt somit ein klassischer Ein-Gruppen-Plan mit Praetest, Posttest und Follow-up vor, der in pädagogisch-psychologischen Studien als vor-experimentelles Design bezeichnet würde (Rost, 2007).

Die Praetests wurden unmittelbar nach der Begrüßung der Schüler/innen und der Erstinformation zum Sinn und den Zielen des Projektes durchgeführt. Alle beteiligten Klassen, mit Ausnahme einer, hatten bis dahin im Regelunterricht nichts zu den jeweiligen Teilgebieten (Elektrizitätslehre und Optik) durchgenommen. Die Praetests sollten somit das vorunterrichtliche Wissen abbilden.

Die Posttests wurden unmittelbar nach der Intervention (Tutoring) durchgeführt und sollten somit den nach der Intervention vorhandenen Wissensstand beschreiben. Die Durchführung sowohl der Prae-, als auch der Posttests wurde jeweils von der Autorin selbst gleitet, um die Versuchsbedingungen konstant zu halten. Es wurden zu allen Testzeitpunkten dieselben Tests verwendet. Die Testdauer betrug etwa 15 Minuten.

Für die Follow-up Tests wurde ein Zeitpunkt gewählt, der mindestens zwei Wochen nach der Intervention lag. In manchen Fällen wurde aber auch erst fünf Wochen nach der Intervention getestet. Hätte man hier eine Laborstudie betreiben können, wäre es sicherlich besser gewesen, das Zeitintervall für alle Klassen konstant zu halten. Unterschiedliche Zeitabstände bedeuten allerdings nur leicht unterschiedliche Testbedingungen für die Langzeitwirkung von CAPT: Je später der Zeitpunkt des Follow-

up Test nach der Intervention ist, desto strenger ist das Testkriterium (Bortz & Döring, 2003). Damit bleibt eine Interpretierbarkeit trotzdem erhalten. Unter den Umständen der hier durchgeführten Feldstudie war das aber nicht möglich. Organisatorische und schulinterne Gründe (Ferien, Schularbeiten, Projekttage,...) sind dafür zu nennen. Auf Wunsch der Klassenlehrer/innen nach mehr Flexibilität in der Zeiteinteilung wurden die Follow-up Tests in den meisten Fällen in Abwesenheit der Forscherin durchgeführt, was aber nur zu wenig geänderten Testbedingungen führte, da die Lehrer/innen die bei den anderen Tests angewandten Durchführungsbedingungen kannten und fortführten.

Die Follow-up Tests sollen eine Beurteilung zulassen, wie nachhaltig das in der CAPT-Intervention erworbene Wissen den Schüler/innen zur Verfügung steht. Es kann angenommen werden, dass das erworbene Wissen einige Zeit nach der Intervention erst in vollem Umfang zur Verfügung steht. Danach jedoch unterliegt es dem Vergessen, wie das für jede Unterrichtsmethode oder Lernform in unterschiedlichem Ausmaß ebenso zutrifft.

Im Zuge der Planung der Untersuchung wurde eine Entscheidung zu eben diesem Prae-Post-Follow-up – Testdesign getroffen. Das bedeutet aber, dass ein Vergleichsgruppen- oder randomisiertes Kontrollgruppendesign verworfen wurde.

Der Hauptgrund dafür liegt darin, dass es sich hier nicht nur um eine Evaluation, sondern auch um die Entwicklung und Adaptierung einer Intervention für den naturwissenschaftlichen Unterricht handelte. Auch wäre es aus schulpraktischen, personellen und zeitlichen Gründen nicht möglich gewesen, weitere Klassen aufzutreiben, geschweige denn hier eine randomisierte Aufteilung zu ermöglichen.

Ein derartiges Forschungsdesign wäre nach den Richtlinien pädagogisch-psychologischer Studien zu den schwächeren Designs zu zählen, die vorsichtig interpretiert werden müssen (Rost, 2007). Dem kann jedoch entgegen gehalten werden, dass es sich hier um empirische Fachdidaktik handelt, die ein weites Spektrum an Anforderungen zwischen Fachwissenschaft, Methodik, Pädagogik und Didaktik abdecken muss. Fachdidaktik lebt von der Vielfalt von Beiträgen und Blickwinkeln unterschiedlicher Bereiche und weniger von domänenspezifischen Standards. Darüber hinaus existieren zur Lernwirksamkeit unterschiedlicher Unterrichtsstile Studien (z.B. Hattie, 2009), weshalb es möglich ist, die Lernwirksamkeit von CAPT auf Basis von Effektstärken auch ohne ein

Vergleichsgruppendesign mit der des herkömmlichen Unterrichts zu vergleichen. Die Frage nach zusätzlichen Einflüssen, die nicht kontrolliert wurden und die möglicherweise den Unterrichtserfolg bedingt haben, ist jedoch mit großer Sorgfalt zu untersuchen.

Dieses quasiexperimentelle Ein-Gruppen-Design verfügt dennoch über ausreichend interne Validität, da Störvariable, deren Einflüsse von vornherein auf der Hand lagen (z.B. Gollwitzer & Jäger, 2009) von Beginn an kontrolliert wurden. Zu den einflussreichsten Störvariablen zählten das Vorwissen und das zwischenzeitliche Geschehen zwischen den Testzeitpunkten. Das Vorwissen wurde durch eben die Praetests kontrolliert. Um das Geschehen zwischen den Testzeitpunkten zu kontrollieren wurden die beteiligten Lehrer/innen genauestens instruiert, was sie mit den Schüler/innen zwischenzeitlich machen durften, bzw. was zu unterlassen war. Auf keinen Fall sollte es so sein, dass die Schüler/innen quasi „Nachhilfe“ durch ihre Lehrer/innen bekamen. Die Einhaltung dieser Bedingung war Teil der Projektvereinbarungen. Im Gegenzug wurde den Lehrpersonen Anonymität bezüglich der Klassen- oder Schulergebnisse zugesichert, da der Fokus auf der Methode und nicht in einem Leistungsvergleich der Standorte lag.

Um Untersuchungsleiterartefakte gering zu halten, wurden alle Instruktionen nach einem genau festgelegten Vorgehen durchgeführt, nach Möglichkeit auch von derselben Person. Abweichungen vom Versuchsablauf wurden dokumentiert und die Abfolge wurde konstant gehalten (nach: Bortz & Döring, 2003).

Als potenzielle zusätzliche Störvariable könnte von einem Reifungseffekt ausgegangen werden, der aber wegen der kurz aufeinander folgenden Testzeitpunkte nicht zu erwarten war. Einflüsse der Drop-out Rate zu den einzelnen Testzeitpunkten wurden durch einen sorgsamen Umgang mit fehlenden Werten minimiert. A posteriori wurden die Ergebnisse des multiplen linearen Regressionsmodells (vgl. Kapitel 7.1.3) herangezogen, um die interne Validität abzusichern. Um den Ergebnissen vorzugreifen, konnten Aussagen darüber getroffen werden, welche der kontrollierten Parameter signifikanten und praktisch bedeutsamen Einfluss hatten und welche nicht.

4.2. Zur Herkunft der Schüler/innen und Stichprobenziehung

Stichprobe und Schulstandorte

Das Sample, das für diese Untersuchungen zur Verfügung stand, bestand zu Beginn der Studie aus N = 172 Schüler/innen, die sich in 9 Schulklassen zu vier Schulstandorten mit variierenden Schulformen aufteilten (Tabelle 4.1).

Klasse	Schulstufe[8]	Schulform	Schulstandort
1	8	Hauptschule	1
2	6	Hauptschule	2
3	6	Hauptschule	2
4	7	Hauptschule	3
6	7	AHS	4
7	6	Hauptschule	4
8	6	Hauptschule	4
9	7	Hauptschule	4
10	7	Hauptschule	4

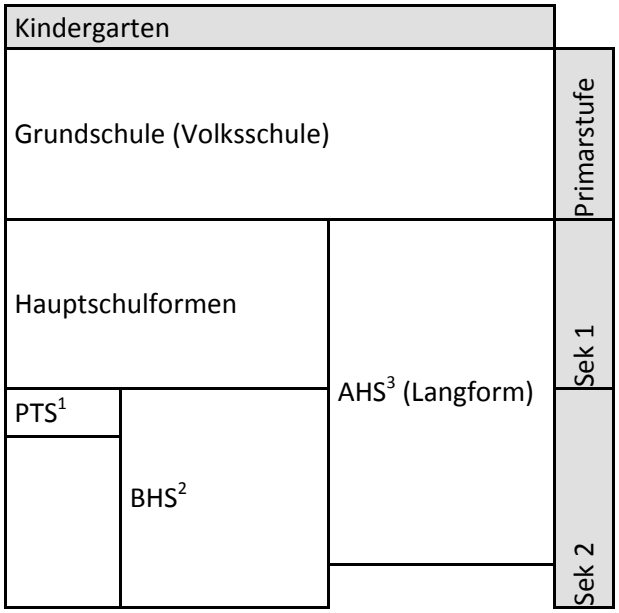
Tabelle 4.1: Überblick über die Schulstufen und Schularten an den fünf Schulstandorten.

Einen groben Überblick über das österreichische Schulwesen gibt Abbildung 4.1. Dabei bilden leicht variierende Hauptschulformen und die Unterstufe der Allgemein bildenden höheren Schule (AHS) die hier behandelte Sekundarstufe 1, entsprechend den Schulstufen fünf bis acht und einem Alter der Schüler/innen von 10 bis 14 Jahre.

AHS und Hauptschulformen unterscheiden sich sowohl hinsichtlich der Klientel als auch der Zielsetzungen: Nach einem verpflichtenden Kindergartenjahr wechseln die Kinder im Alter von sechs Jahren in die Grundschule (Volksschule), die vier Schulstufen hat. Danach werden die Abgänger/innen im Alter von 10 Jahren entsprechend ihrer kognitiven, sprachlichen und sozialen Fähigkeiten, bzw. der Prognosen darüber auf die weiter führenden Schulen aufgeteilt. Die allgemein bildende höhere Schule ist für leistungsstärkere Jugendliche gedacht, die Reifeprüfung und danach ein Hochschulstudium anstreben. Die Hauptschule bietet Raum für Jugendliche, die nach der Volksschule sprachlich oder kognitiv in ihrer Entwicklung weniger weit sind, mit der

[8] Schulstufe zu Beginn der Datenerhebung, thematisch der Elektrizitätslehre zuzuordnen

Möglichkeit, sie besonders zu fördern. De facto drängen jedoch in den Ballungsgebieten, wie im hier untersuchten Raum um Wien, bis zu drei Viertel der Grundschulabgänger/innen in die AHS (Statistik-Austria, 2014). In ländlichen Räumen ist dieses Phänomen aufgrund weiter Schulwege und/oder fehlender Infrastruktur weniger stark ausgeprägt. Mit ein Grund für das verstärkte Streben in die AHS liegt darin, dass das österreichische Schulsystem zwar auf dem Papier Durchlässigkeit verspricht, es aber in der Praxis gerade im urbanen Raum nur mit erheblichem Aufwand möglich ist, nach der Hauptschule in einer weiter führenden Schule Fuß zu fassen oder in eine Oberstufe einer AHS umzusteigen, obwohl für die AHS-Unterstufe und die Hauptschule dieselben Lehrpläne gelten und dieselben Schulbücher approbiert sind.



[1] Polytechnische Schule; [2] Berufsbildende höhere Schule; [3] Allgemein bildende höhere Schule

Abbildung 4.1: Schulformen in Österreich im Überblick

In den Hauptschulen, wie in anderen Formen der Pflichtschule[9], finden sich vermehrt Risikoschüler/innen. Deren Anteil beträgt im Bereich der Naturwissenschaften 37 %, im Bereich Lesen sogar 48 % (OECD, 2010; Schwantner, Toferer, & Schreiner, 2013), wobei Schüler/innen mit Migrationshintergrund, wie sie laut OECD definiert sind, in der Gruppe der Risikoschüler/innen in allen PISA 2006 Testländern deutlich stärker vertreten sind (Schmirch, 2009, S. 107). Um hier gezielte Fördermaßnahmen zu implementieren und

[9] Die allgemeine Pflichtschule (APS) umfasst Sonderschule, Hauptschulformen und Polytechnische Schulen. Im Rahmen der oben zitierten Publikationen zu PISA 2006 sind die Daten für die Hauptschule allein nicht verfügbar.

langfristig das Hauptschulniveau dem der AHS anzugleichen, wurden in den letzten Jahren unterschiedliche Formen der Hauptschule kreiert, die die Namen Kooperative Mittelschule (KMS), Neue Mittelschule (NMS) oder Wiener Mittelschule (WMS) tragen. Diese Formen waren kürzlich Gegenstand umfangreicher Prüfungen, wobei der österreichische Rechnungshof die hohen Kosten (Rechnungshof, 2013) und der Nationale Bildungsbericht 2012 die geringe Wirksamkeit der Maßnahmen (Bruneforth, Herzog-Punzenberger, & Lassnigg, 2013) kritisiert. Ob weitere Evaluierungen dieser Schulformen erfolgen werden, ist unklar, nachdem das derzeitige Bildungsministerium alle weiteren Testungen ausgesetzt hat (2014).

In Ermangelung geeigneter Infrastruktur zu empirischer fachdidaktischer Forschung wurde unter den gegebenen Umständen versucht, dennoch ein möglichst breit gestreutes und unverzerrtes Sample zu rekrutieren. Die Klassen mussten selbst organisiert werden, was oft auf Basis von Mundpropaganda passierte. Dabei wurde darauf geachtet, Schüler/innen mit unterschiedlichen sozio-ökonomischen Hintergründen in die Stichprobe aufzunehmen.

Schule 1 ist eine private Hauptschule in einem der „besseren“ Bezirke Wiens. Das bedeutet, dass der durchschnittliche sozio-ökonomische Status der Eltern höher und auch der Anteil an Schüler/innen mit nicht-deutscher Muttersprache geringer ist. Darüber hinaus ist diese Schule eine Praxisschule einer Lehrerbildungseinrichtung[10], die Pflichtschullehrkräfte ausbildet. Die dortige Physiklehrerin ist eine ausbildende Lehrkraft für Studierende, die möglicherweise aufgrund dieser Tatsache über ein höheres fachdidaktisches Wissen verfügt, was wiederum ihren Unterrichtsstil beeinflussen könnte.

Schule 2 ist ebenfalls eine Hauptschule, allerdings eine öffentliche mit bis zu 90 % Schüler/innen mit nicht-deutscher Muttersprache und laut Auskunft der Klassenlehrer/innen mit einem äußerst niedrigen durchschnittlichen sozio-ökonomischen Status der Eltern. Die dort arbeitenden Lehrer/innen waren ebenfalls in der Lehrer/innenbildung tätig.

[10] Diese nennt sich „Pädagogische Hochschule“, ist aber aus unterschiedlichen Gründen nicht mit der deutschen Einrichtung gleichen Namens zu vergleichen. Hier werden in Österreich z.B. Hauptschullehrer/innen ausgebildet.

Schule 3 ist wiederum eine private Praxisschule derselben Lehrerbildungseinrichtung wie Schule 1, jedoch in der Peripherie Wiens angesiedelt. Das hat zur Folge, dass nicht nur der sozio-ökonomische Status höher ist, sondern auch, dass es kaum Schüler/innen mit nicht-deutscher Muttersprache gibt.

Schule 4 ist ein Schulzentrum etwa 30 km außerhalb von Wien, das in privater, kirchennaher Trägerschaft ist. Hier fanden sich keine Praxislehrer/innen, dafür aber jene Schüler/innen mit höherem sozio-ökonomischen Status, da hier Schulgeld bezahlt werden muss.

Migrationshintergrund und Muttersprachen

CAPT als Unterrichtsmethode basiert zu einem nicht unerheblichen Teil auf verbaler Interaktion, sowohl auf der passiven Ebene des Verstehens, wie auch auf der aktiven des Erklärens. In der Bildungsforschung existiert ein Konsens darüber, geringe Kenntnisse in der Unterrichtssprache als Ursache für geringere Lernerfolge zu sehen (Eckhardt, 2008). Um eine Erklärung für mögliche Leistungsunterschiede zwischen Schüler/innen einheimischer Herkunft und solchen mit Migrationshintergrund geben zu können, wurde für die teilnehmenden Schüler/innen erhoben, ob sie Deutsch oder eine andere Sprache zu Hause sprechen. Das entspricht zwar nicht exakt einer Erhebung des Migrationshintergrundes erster und zweiter Generation nach der Definition von PISA (Breit, 2009), stellt aber im Rahmen der Untersuchungen die entscheidende Variable dar. Mit ihr wird versucht, zumindest ein elementares Maß für die Sprachbeherrschung der Schüler/innen zu finden. Schüler/innen mit Migrationshintergrund und somit in den meisten Fällen mit nicht-deutscher Muttersprache besuchen häufiger eine Pflichtschule, wie die hier genannte Hauptschule. Für ganz Österreich sind das in Zahlen 14 % der Einheimischen, aber 27 % der Migranten. Bei den Jungen ist das Verhältnis noch extremer: 32 % der männlichen Migranten besuchen die Pflichtschule, aber nur 18 % der männlichen Einheimischen. Wie oben beschrieben, stammt ein großer Teil der Schüler/innen des Samples aus der Hauptschule, einem Teilbereich der Pflichtschule. In Ballungsräumen wie dem untersuchten Bereich um die Bundeshauptstadt Wien ist aufgrund des höheren Anteils an Migranten insgesamt mit einem noch größeren Prozentsatz an Schüler/innen mit Migrationshintergrund in den einzelnen Klassen zu

rechnen. Die genauen Zahlen dazu finden sich in den Ergebnissen (Kap. 7.1.1 für die Elektrizitätslehre und Kap. 7.2.1 für die Optik).

Nun stellt jedoch Migrationshintergrund einen eindeutigen Nachteil im Kompetenzerwerb dar, nicht nur für die Lesekompetenz, sondern, vielleicht entgegen anderslautender Vermutungen, auch für den Erwerb aller anderen Kompetenzen. Ungenügende Beherrschung der Unterrichtssprache kann nicht durch Stärken in anderen Bereichen kompensiert werden (Baumert & Schümer, 2001). Darüber hinaus legen die Autoren dar, dass fast jedes systematische und selbstständige Lernen sprachenbasiert ist und somit ist zu vermuten ist, dass auch für die CAPT-Intervention die Sprachbeherrschung eine nicht unerhebliche Rolle spielt. Da sich aufgrund der PISA-Daten herausstellte, dass weder die kulturelle Distanz, noch der sozioökonomische Hintergrund für die Unterschiede in der Bildungsbeteiligung und Kompetenzerwerb von migrantischen Jugendlichen im Vergleich zu einheimischen relevant ist (ebd.), wurde die Variable *Muttersprache* erhoben. Die Beherrschung oder Nichtbeherrschung der Unterrichtssprache kann sich im Rahmen der CAPT-Intervention vielfach auswirken: Zum einen sollen die Schüler/innen in der Lage sein, der Instruktion zu folgen oder sogar selbst eine zu gestalten. Zum anderen sind begleitend zur Intervention eine Vielzahl an Fragebögen auszufüllen, für deren Beantwortung die Inhalte der einzelnen Items erfasst werden müssen.

Zusammenfassend kann zur Auswahl der Schulen festgestellt werden, dass es betreffend des sozioökonomischen Status der Schüler/innen eine breite Streuung gab, wie auch betreffend der Muttersprachen (Details dazu siehe Kapitel 7.2.1). Die Hauptschüler/innen sind hier deutlich in der Überzahl, was aber mit den Zielen der Studie zusammenpasst, da CAPT als universell einsetzbare Unterrichtsmethode beforscht werden sollte und nicht als Maßnahme zur Förderung begabter Schüler/innen. Die Mischung zwischen privaten und öffentlichen Schulen sollte insofern zu keiner Verzerrung im Sample führen, da in Österreich für alle diese Schulen dieselben Lehrpläne gelten und gemäß dem Konkordat auch die Lehrer/innen „lebende

Subventionen“ des Staates an die Schulen sind[11]. Insbesondere unterliegen alle dem gleichen Dienstrecht und erhalten die gleiche Bezahlung. Hinsichtlich des sozio-ökonomischen Status muss aber wohl davon ausgegangen werden, dass sich die beforschten Klassen unterschieden. Das ist einer der Gründe, warum in den Kapiteln 7.1.2 und 7.2.1 die Praetests einer genauen Analyse unterzogen wurden.

Klumpenstichproben

Die vorliegende Art der Stichprobe, bei der ganze Klassen ausgewählt und getestet werden, bezeichnet man als Klumpenstichprobe. Von einem Klumpen spricht man dann, wenn sich eine Grundgesamtheit in disjunkte Teilgesamtheiten zerlegt. Im Falle der vorliegenden Untersuchung ist es unumgänglich mit geklumpten Stichproben zu arbeiten, da auch die Tauglichkeit von CAPT für den Schulalltag und damit für ganze Klassen zu prüfen war.

Der Vorteil von Klumpenstichproben liegt darin, dass man keine vollständige Liste der Elemente der Grundgesamtheit benötigt und daher die Kosten einer Untersuchung oft geringer sind, jedenfalls aber der Aufwand. Im vorliegenden Fall wurde die Durchführung der Studie erst durch eine Klumpenziehung gewährleistet. Darüber hinaus war es ganz im Sinne der Untersuchung, CAPT als eine in der Schulpraxis leicht zu implementierende Arbeitsform, die mit ganzen Klassen durchgeführt werden kann, zu testen und zu beurteilen.

Der Nachteil an Klumpenstichproben liegt vor allem darin, dass die Schätzungen der Populationsparameter im allgemeinen höhere Standardfehler aufweisen, als bei einem streng randomisierten Sample (Apostolopoulos & Schulz, 2003). Dieser Umstand wird als *Designeffekt* bezeichnet. Die Stärke dieses Effektes hängt vom Maß der Homogenität innerhalb der Klumpen und ihrer Zahl ab. Ist die Intraklassenkorrelation hoch, ist auch das Ausmaß der Verklumpung hoch. In diesem Fall ist eine hohe Anzahl an Klumpen (hier: Klassen) für Signifikanztestungen nötig. Es ist günstig, wenn aus einem Klumpen alle Elemente gezogen werden (*one stage cluster sampling*), was im vorliegenden Fall auch getan wurde. Die Ursache für einen Designeffekt liegt darin begründet, dass

[11] Das Konkordat ist ein Vertrag zwischen Vatikan und Einzelstaaten wie Österreich, in dem u.a. das Eherecht, der katholische Religionsunterricht an Schulen und die Finanzierung der katholischen Privatschulen geregelt ist. Für letztere ist vereinbart, dass der Staat für die Personalkosten aufkommt.

spezifische Eigenschaften der Elemente der Grundgesamtheit (Schüler/innen) die Zugehörigkeit zu einem bestimmten Klumpen (Klasse) bestimmen. Die Schüler/innen einer Klasse sind einander somit ähnlicher als die Schüler/innen verschiedener Klassen, da die Verteilung der Schüler/innen auf die Klassen durch die Eigenschaften der Schüler/innen mitbestimmt ist und daher nicht ganz zufällig. Analysen, die auf einer zufälligen Stichprobenziehung beruhen, sind somit mit Vorsicht zu interpretieren.

Rost (2007) stellt unterschiedliche Vorgangsweisen zur Handhabung von Klumpenstichproben vor. Man kann die Variablen vor der Analyse standardisieren (Transformation auf eine N(0;1)-Verteilung). Oder man berechnet Korrelationen nach Klassen getrennt zu und mittelt dann über die Klassen. Eine weitere Möglichkeit besteht darin, aus jeder Klasse einen Schüler zufällig zu ziehen. Damit hätte man eine Zufallsstichprobe realisiert, benötigt aber sehr viele Klumpen. Eine weitere Option stellt eine Analyse auf Klassenebene dar.

Im Fall der vorliegenden Untersuchungen wurde im Bereich der Elektrizitätslehre auf eine besondere Berücksichtigung der Klumpenproblematik verzichtet, da sich das Sample in den Praetests als sehr homogen herausstellte. Für die Optik wurde aufgrund der stark differierenden Praetests eine Analyse auf Klassenebene durchgeführt, um der Klumpenstichprobe gerecht zu werden.

4.3. Inhaltliche Fokussierung

Eine inhaltliche Fokussierung ist aus mehreren Gründen sinnvoll. Zum einen plädieren einige Lernforscher *(learning scientists*) für die Entwicklung gegenstandsspezifischer, lokaler Theorien (z.B. Barab & Squire, 2004; Prediger & Link, 2012), aus der Erfahrung heraus, dass spezifische Kontexte zu spezifischen Lernschwierigkeiten führen, für die es nur lokale, aber keine globalen Lösungen gibt.

Zum anderen empfiehlt Zinn (2009) in seiner Conclusio für zukünftige Arbeiten eine feste inhaltliche Fokussierung, um zu erwartende Lernerfolge genauer beschreiben zu können, nachdem er selbst mit unterschiedlichen Lerninhalten gearbeitet hatte.

Die Auswahl spezieller Themen, die im Rahmen der CAPT-Intervention behandelt wurden, orientierte sich stark an den bestehenden Curricula der Sekundarstufe 1 (BMUKK, 2000). Was die Tutoringprozesse mit Vor- und Grundschulkindern betrifft, ist eine Behandlung im Rahmen des Vor- und Grundschulunterrichtes zwar möglich, wenn auch nicht explizit in den Curricula gefordert. Auf Basis der Curricula einerseits und der Literaturrecherche zu Schülervorstellungen andererseits entschied sich das Projektteam für Elektrizitätslehre und Optik, da sich für beide Bereiche Bezüge zu Vorarbeiten herstellen ließen und die curriculare Validität gegeben ist.

Für die Beantwortung der Forschungsfragen hat sich während des ersten Studienjahres herausgestellt, dass es nicht möglich war, allen Forschungsfragen in einem einzigen Jahr nachzugehen und somit die Untersuchungen zwei Mal parallel durchzuführen. Das lag vor allem an der Schwierigkeit Themen zu finden, die in *allen* beforschten Schulstufen auf Basis der Curricula möglich sind. Nach dem ersten Studienjahr, und nachdem die Forschungsfragen 1 bis 5 und 7 ausreichend beantwortet waren, wurden für das zweite Studienjahr die Forschungsfragen 1 und 6 zur Beantwortung ausgewählt. Darüber hinaus erschien es aufgrund der qualitativ stark unterschiedlichen Wissenstests, die für das Teilgebiet der Elektrizitätslehre einerseits und Optik andererseits zur Verfügung standen, sinnvoll, im zweiten Studienjahr andere Forschungsfragen in den Mittelpunkt zu stellen. Außerdem war es ein implizites Ziel der Studie, CAPT anhand verschiedener Themen der Physik innerhalb der zwei Studienjahre zu untersuchen, um die Tauglichkeit möglichst viele Themenbereiche der Physik für diese Methode zu erproben.

Das bedeutete somit, dass im ersten Studienjahr der Fokus auf anderen Dingen lag als im zweiten: Wurde im ersten Jahr anhand der Elektrizitätslehre die Wirksamkeit der Methode an sich und der Einfluss der Rollen untersucht, so wurde der Schwerpunkt im zweiten Jahr darauf gesetzt, zu erforschen, ob CAPT auch bezüglich anderer physikalischer Themen (in diesem Fall Schatten und Spiegel) Effekte zeigt.

Für beide Bereiche, Elektrizitätslehre und Optik, wurde versucht sich inhaltlich an den durch die Literatur bestätigten Basiskonzepten zu orientieren. Als Basiskonzepte werden Konzepte bezeichnet, die einerseits hinsichtlich der didaktischen Rekonstruktion die Grundlage für weitere Konzepte bilden. Andererseits sind sie in den Curricula erstgereiht. Nicht zuletzt findet man bereits in der Grundschuldidaktik (z.B. Wiesner,

2004a, 2004c) diese Konzepte an erster Stelle. Da für die Effektivität von Tutoring-Programmen eine gute Strukturierung förderlich ist (P. A. Cohen, et al., 1982; Topping, 2005), wurden die Themen, basierend auf den dazugehörigen Basiskonzepten, in sachlogischer Reihenfolge zur Bearbeitung im Tutoringprozess ausgewählt.

Elektrizitätslehre

Im Bereich der Elektrizitätslehre wurden die Interventionen an Basiskonzepten orientiert, die als Grundlage für ein elaboriertes Verständnis notwendig sind (Duit & Rhöneck, 1998; Shipstone, 1984). Diese Basiskonzepte sind:

- Elektrische Stromkreise müssen geschlossen sein, damit Strom fließen kann.
- Der elektrische Strom hat eine bestimmte Richtung im Stromkreis.
- Die Stromstärke ist überall im Stromkreis konstant. Insbesondere „verbraucht" ein Gerät keinen Strom.
- Es gibt einen Zusammenhang zwischen Stromstärke, elektrischem Widerstand und der Helligkeit gleichartiger Glühlämpchen.
- Überlegungen zur Gesamtstromstärke in einer Parallelschaltung.

Optik

Innerhalb der Optik konnten zwei Teilbereiche, die Themenkreise zu Schatten und Spiegel, identifiziert werden, besonders geeignet erschienen die unterschiedlichen Altersgruppen der Tutees zu bedienen. Als Basiskonzepte der Intervention in der Optik wurden daher – in Übereinstimmung mit der Forschungsliteratur zu Schülervorstellungen (Andersson & Kärrquist, 1983; Colin, Chauvet, & Viennot, 2002; Galili & Hazan, 2000; Goldberg & McDermott, 1986; Guesne, 1984, 1985; Wiesner, 1992, 2004b, 2004c) – die folgenden Bereiche festgelegt:

- Vermittlung einer korrekten Sehvorstellung
- Lichtausbreitung als Strömen des Lichtes
- Schatten als Lichtmangel
- Schatten als Raumbereich im Unterschied zum Schlagschatten
- Ort des Spiegelbildes hinter dem Spiegel
- Der Spiegel vertauscht Vorder- mit Rückseite, nicht linke mit rechter Seite.
- Zusammenhang zwischen Reflexionsgesetz und Spiegelbild

Während die Tutoren in den meisten Fällen aus der Sekundarstufe 1 stammten, waren die Tutees einerseits jüngere Schüler/innen aus dieser Stufe, andererseits Volksschüler/innen oder sogar Kindergartenkinder (Vorschulgruppen). Um an bestehende Curricula anschließen zu können, wurden deshalb die Einheiten für die jüngeren Tutees (aus der Volksschule oder dem Kindergarten) zum Thema Schatten gestaltet, die für die älteren (aus der Sek 1) zum Thema Spiegel (vgl. Tabelle 7.23). Den Ausgangspunkt aller Einheiten bildete immer die Vermittlung einer korrekten Sehvorstellung.

Design der Lernmaterialien und Planung der Intervention

Das Design der Lernmaterialien und der Interventionen selbst, egal ob es das Mentoring oder das Tutoring betraf, richtete sich nach Ergebnissen aus der Forschung über Schülervorstellungen und den daraus identifizierten grundlegenden Konzepten, die vorrangig adressiert wurden. Dazu wurden entsprechend ausgearbeitete und erprobte Materialien (z.B. Haagen-Schützenhöfer & S., 2014) adaptiert. Dabei wurden empfohlene Strategien der Instruktion umgesetzt (Shaffer & McDermott, 1992), nicht zuletzt solche, die konstruktivistischen Lerntheorien entsprangen (Widodo & Duit, 2004). Beispielmaterialien zum Mentoring und zum Tutoring sind auszugsweise in den Anhängen 13.5 und 13.6 zu finden.

Da die Klassen nicht alle exakt derselben Altersstufe angehörten und ihre Tutees ebenfalls nicht genau gleich alt waren, ergaben sich in Abstimmung darauf für jede Klasse leicht variierende Unterrichtsmaterialien. Diese Materialien waren auf die zu adressierenden Schülervorstellungen abgestimmt.

Ein wichtiges Faktum bezüglich der CAPT-Intervention war es, dass mit allen beteiligten Lehrkräften vereinbart wurde, die Inhalte des CAPT nicht im regulären Unterricht zu thematisieren. Das betraf einerseits die Unterrichtszeit vor der CAPT-Intervention: Obwohl CAPT nicht an allen Schulen zur selben Zeit durchgeführt wurde, sollten die Schüler/innen bis dahin nichts über die Inhalte erfahren.

Zudem wurde mit den Lehrkräften auch vereinbart, dass zwischen den einzelnen Interventionen und Testzeitpunkten (Praetest – Posttest – Follow-up-Test) keine inhaltliche Arbeit zu den Themen der CAPT-Intervention stattfand und keine Fragen aus den Tests beantwortet wurden. Nach besten Wissen und Gewissen und dem Ehrenwort

der Beteiligten wurde das auch wie versprochen durchgehalten (vgl. S. 52 Überlegungen zur internen Validität der Studie).

4.4. Mentoring und Tutoring

Die eigentliche inhaltliche CAPT-Intervention, abgesehen von den durchzuführenden Tests, gliederte sich in zwei Teile: das Mentoring und das Tutoring. Das Mentoring erfolgte zeitlich vor dem Tutoring und beinhaltete eine Art Einführung für die Klassen der zukünftigen Tutor/innen. Das Tutoring bezeichnet die eigentliche Peer-Instruktion.

4.4.1. Ziele des Mentoring

Entsprechend der Empfehlungen von Fogarty und Wang (1982) ist es sinnvoll, die zukünftigen Tutor/innen einem Training zu unterziehen, wenn, wie in diesem Fall, die Inhalte zumindest teilweise neu sind. Während Fogarty und Wang diese Empfehlung für Mathematik-Nachhilfeprogramme und Programme zur Verbesserung der Grundfertigkeiten am Computer formulieren, scheint ein Training im Falle der Physik, wo ein Konzeptwechsel angestrebt wird, umso sinnvoller zu sein. Dieses Training wird in der Folge als *Mentoring* bezeichnet. Zur Begriffsklärung sei erwähnt, dass es sich hier um eine Instruktion handelt, die einem an fachdidaktischer Forschung geleiteten Unterrichtssetting entspricht. *Mentoring* ist daher nicht im Sinne der ersten Phase des Lehrlings-Meister-Modells (*cognitive apprenticeship*) gemeint, wo der Meister (Mentor) dem Lehrling etwas vorzeigt (vgl. Topping, 2005).

Das Mentoring hat zwei Hauptaufgaben zu erfüllen: es soll die zukünftigen Tutor/innen mit den Inhalten vertraut machen, daher zunächst im Sinne einer *explicit instruction* (Clark, et al., 2012) zu einer fachlichen Klärung beitragen. Auf der anderen Seite sollen die Tutor/innen mit ihren eigenen Vorstellungen konfrontiert werden und die Gelegenheit erhalten, diese zu reflektieren. Es könnte sein, dass sich bei ihnen die eine oder andere Fehlvorstellung verbirgt, die eine Entwicklung adäquater Vorstellungen behindert. In diesem Fall bietet das Mentoring die Gelegenheit, die eigenen Vorstellungen in Richtung wissenschaftlich korrekter Vorstellungen weiterzuentwickeln. Dieser Lern- und Reflexionsprozess soll auch dazu beitragen, dass die Tutoren sensibel für die Vorstellungen ihrer Tutees werden und Wege finden können, sie in geeigneter

Weise in Richtung wissenschaftlich anschlussfähiger Vorstellungen zu verändern. Das Mentoring versteht sich als Möglichkeit, wenn nicht sogar als Aufforderung, Konstruktionsprozesse in Gang zu setzen, und bildet daher eine Lernumgebung im Sinne konstruktivistischer Lerntheorien.

Ein weiterer Aspekt des Mentorings war es, für das folgende Tutoring Problemstellungen und Lernmaterialien auszusuchen oder sogar zu erfinden, die für die jeweilige Tutee-Klasse geeignet erschienen. Hier wurde den zukünftigen Tutoren in einem gewissen Rahmen die Wahl gelassen, welche Lernmaterialien sie, in Abhängigkeit des Alters der Zielgruppe, auf welche Weise einsetzen wollten.

Dazu wurde innerhalb der Tutorenklasse unter Anleitung der Forscherin zunächst eine Einigung darüber erzielt, welche Konzepte überhaupt angesprochen werden sollten. Danach wurden geeignete Experimente und Aufgaben für die Umsetzung gewählt. Die Tutoren wurden somit als Expert/innen für das Lernen der Jüngeren angesprochen, was sie nach dem Verständnis von CAPT auch sind. Das wiederum sollte zur Folge haben, dass im Sinne der SDT ihre *perceived competence* und ihre *autonomy* angesprochen wird und sie so (intrinsisch) motiviert werden.

Alle Mentorings wurden von derselben Person nach demselben Muster durchgeführt, was zumindest diesen Faktor *konstant* halten sollte. Wenn schon nicht der Einfluss des Mentorings selbst untersucht wurde, sollte somit zumindest sichergestellt sein, dass alle Klassen demselben Einfluss unterlagen.

4.4.2. Praktische Durchführung des Mentoring

Zu Beginn wurden alle Schüler/innen begrüßt und ihnen kurz das Projekt sowie die Unterrichtsmethode vorgestellt. Es wurde die Rolle der Schüler/innen als Expert/innen für das Lernen der Jüngeren betont. Danach wurden die Einstiegstests administriert.

Nach dem Einstiegstest erhielten die Schüler/innen eine Liste, die theoretische und experimentelle Aufgaben enthielt. Zusätzlich wurde in Klassenstärke experimentelles Material ausgegeben. Die Schüler/innen wurden dazu angehalten, einzeln oder in Partnerarbeit mit dem Sitznachbarn diese Aufgaben zu behandeln. Musterbeispiele dieser Aufgabenteile sind im Anhang 13.5 nachzulesen. Wenn möglich, sollten

theoretische Vorhersagen experimentell überprüft werden und danach Erklärungen gefunden werden.

Diese Vorgangsweise entspricht jenem Konzept, das es am ehesten ermöglicht, Experimente sinnvoll in den Unterricht einzubinden. Es ist die die P-O-E – Strategie (White & Gunstone, 1992): *predict – observe – explain*. Bevor Schüler/innen zu experimentieren beginnen, werden Vorhersagen über den möglichen Versuchsausgang gemacht. Im Experiment wird dann überprüft, ob diese Vorhersage zutrifft. Zum Schluss wird versucht eine Erklärung für den Versuchsausgang zu finden.

Auf diese Phase des selbstgesteuerten Lernens folgte eine Diskussion in der Gruppe, die von der Forscherin moderiert wurde. Es wurden auftretende Fehlvorstellungen den korrekten Erklärungsmustern gegenübergestellt und versucht, Gründe für die Fehlvorstellungen zu finden. Diese Diskussion im Sinne konstruktivistischer Lerntheorien wird als fruchtbare Basis für die Weiterentwicklung der Konzepte der Tutoren angesehen (Widodo & Duit, 2004).

In der folgenden Phase wurden die Schüler/innen dazu angehalten, jene Aufgaben und Experimente auszuwählen, die für ihre spezielle Tuteegruppe am geeignetsten erschienen.

Dabei wurde auch die Möglichkeit eröffnet, eigene Aufgaben zu erfinden. Diese Möglichkeit wurde in wenigen Fällen genutzt. Ein besonders kreatives Beispiel dafür ist eine Schülergruppe aus Klasse 4, die im Bereich der Optik, konkret zu Licht und Schatten, weitere Aufgaben für ihre Grundschul-Tutees entwickelten. Sie bastelten aus farbigen Hefteinbänden Folien für ihre Taschenlampen und erzeugten so farbige Schatten.

Die Dauer dieses Mentoring erstreckte sich zwischen 60 und 80 Minuten, fand mit oder ohne Pause statt, je nach den zeitlichen Ressourcen an den einzelnen Schulen. Unmittelbar vor dem Tutoring erfolgte eine ausführliche Wiederholung der Inhalte des Mentorings. Mehr dazu wird später ausführlicher erklärt werden. Das Mentoring wurde zwischen eineinhalb und zweieinhalb Wochen (in einem Ausnahmefall vier Wochen) vor dem eigentlichen Tutoring durchgeführt.

4.4.3. Vorbereitung und Charakteristik des Tutorings

Nachdem zwischen Mentoring und Tutoring einige Zeit vergangen war, erschien es sinnvoll, die Tutor/innen die einzelnen Aufgaben und Materialien dazu in aller Ruhe nochmals rekapitulieren zu lassen. Daher wurden weitere 20 bis 30 Minuten als Vorbereitungszeit veranschlagt. Aufgabe war es, die im Mentoring ausgesuchten Fragen noch einmal durchzugehen, die experimentellen Materialien dazu herzurichten, zu prüfen und die Experimente noch einmal durchzuführen. Sinnvoller Weise sollte das in Partnerarbeit in Form eines kleinen Rollenspiels geschehen.

Da es sich bei der Mehrzahl der Schüler/innen um kognitiv nicht sehr leistungsfähige Jugendliche handelte, wurden in Absprache mit den Klassenlehrer/innen für jede Aufgabe *cue card*s, Hilfskärtchen, angefertigt und an die Schüler/innen verteilt. Auf der Vorderseite befanden sich die Fragen oder Aufgaben, anhand derer das Tutoring stattfinden sollte. Auf der Rückseite befanden sich Lösungshinweise, für den Fall, dass ein Tutor die korrekte Lösung vergessen haben sollte. Diese *cue cards* dienten dazu, den Tutoringprozess zu begleiten (*scaffolding*) und ihn zu strukturieren. Das schien deshalb notwendig, da Erkenntnisse aus der Psychologie zeigen, dass leistungsschwache Schüler/innen, die vielleicht auch noch wenig Vorkenntnisse besitzen, mit hochstrukturierten Unterrichtsmaterialien besser zurechtkommen (Cronbach & Snow, 1981; Helmke & Weinert, 1997). Darüber hinaus stellte Topping (2005) fest, dass Peer-Tutoring eine bessere Wirksamkeit zeigt, wenn es strukturiert abläuft. Genau diese strukturelle Hilfestellung sollten die Kärtchen geben. Ein Musterbeispiel findet sich im Anhang auf Seite 254, wo die Aufgabe und darunter die Lösung angegeben ist.

Den Abschluss der Vorbereitung bildeten Hinweise, wie die Tutoren lehren sollten: Zuerst kam eine Erinnerung an elementare Regeln der Höflichkeit, einander zu begrüßen und sich vorzustellen. Des Weiteren wurde an die Tutoren dringend appelliert, nicht alles selbst zu machen, sondern ihre Tutees die Experimente selbst ausführen zu lassen. Im Falle theoretischer Aufgaben sollten die Tutees ermutigt werden, zuerst selbst Vermutungen aufzustellen. Jedenfalls wurde noch einmal an die P-O-E – Strategie (White & Gunstone, 1992) erinnert, derer sich die Tutoren bedienen sollten.

Das eigentliche Tutoring begann mit dem Zusammentreffen der beiden Klassen und der Bildung von Tutor-Tutee-Dyaden. Dieses Setting wurde aufgrund der Möglichkeit einer

intensiveren Auseinandersetzung, bzw. aufgrund der eingeschränkten Möglichkeit nach Ablenkung, bevorzugt (Topping, 2005). Die Auswahl der Partner erfolgte in den meisten Fällen zufällig, da die Klassen einander nicht gut kannten. In seltenen Fällen bildeten sich Freundespaare oder Paare gleicher Muttersprache. Insbesondere wurde kein Wert auf gleichgeschlechtliche Paare gelegt. Daten zur geschlechtlichen Zusammensetzung der Dyaden wurden nicht erhoben, da dies nicht im Fokus der Forschungsfragen stand. In Fällen, wo die Klassenschülerzahl ungleich war, übernahm ein Tutor zwei Tutees, oder umgekehrt, so dass Trios entstanden.

Die Altersstufe der Tutoren variierte zwischen 12 und 14 Jahre, entsprechend der 6. bis 8. Schulstufe. Die Tutees waren in fast allen Fällen jünger (*cross-age tutoring*), das Alter variierte zwischen 8 bis 13 Jahren, entsprechend den Schulstufen 2 bis 7. Das führte zu einem mittleren Altersunterschied von 2,82 Jahren und einer Standardabweichung von 1,99. In zwei Fällen fand das Tutoring auf gleicher Altersstufe statt (*same-age tutoring*), der Altersunterschied betrug daher Null Jahre. Den Empfehlungen von Robinson et al. (2005) und Fogarty & Wang (1982) folgend sollte der Altersunterschied nicht zu groß sein, da sonst einer der Gelingensfaktoren von Peer-Tutoring, nämlich die sprachliche und soziale Nähe, wegfällt. Diese Bedingung konnte in den meisten Fällen erfüllt werden. Über den Einfluss des Altersunterschiedes ist außer den beiden oben zitierten Arbeiten nicht viel bekannt. Insbesondere zum *same-age tutoring* äußert sich jedoch Hattie (2009) insofern, dass mögliche Effekte schwächer ausfallen. Somit kann gesagt werden, dass die beiden gleichaltrigen Tutoren-Tutee Klassen die Analysen höchstens dahin gehend beeinflussen, dass die beobachteten Effekte möglicherweise zu schwach geschätzt werden. Auszuschließen ist hingegen, dass sie durch diese Konstellation überschätzt werden.

Auch hinsichtlich der Fähigkeiten der Schüler/innen wurde die Paarbildung nicht beeinflusst. So konnte es vorkommen, dass leistungsstarke Tutoren mit leistungsschwachen Tutees zusammenarbeiteten, aber eben auch umgekehrt: dass leistungsschwache Tutoren leistungsstarke Tutees zugeordnet bekamen. Es war nicht in der Intention der Studie, dass nur leistungsstarke Schüler/innen die Tutorenrolle einnehmen sollten. Vielmehr war es gewünscht, dass ganze Klassen mit ihrem gesamten Leistungsspektrum als Tutoren arbeiteten (*class-wide tutoring*). Es wurde hier also auf

cross-ability Basis gearbeitet, entsprechend Toppings Charakterisierung des Prozesses (2005).

4.4.4. Praktische Durchführung des Tutoring

Tutees und Tutoren wurden einander zugeordnet und im Falle großer Gruppen in mehreren Räumen untergebracht. Die Tutoren starteten die Intervention, indem sie ihren Tutees die vorbereiteten Aufgaben zur Bearbeitung gaben und sie anregten, ihre Vermutungen auch anhand von Experimenten zu prüfen. Dieser Prozess sollte zu einer mehr oder minder lebhaften Diskussion führen.

Insgesamt dauerte das Tutoring 30 bis 45 Minuten. Die genaue Dauer war vom Fleiß der Schüler/innen abhängig. Das Ende des Tutorings war gegeben, wenn entweder das Stundenende erreicht war oder die Schüler/innen meldeten, dass sie schon fertig seien und auch nichts mehr zu diskutieren hätten.

Da es an den Schulstandorten unterschiedlich viele Klassen gab, die an CAPT beteiligt waren, wurde entsprechend der zur Verfügung stehenden Anzahl an Klassen das Studiendesign den Umständen angepasst. Das führte dazu, dass einzelne Klassen mehr Interventionen hatten als andere. Und es führte in weiterer Folge dazu, dass die Schüler/innen einzelner Klasen in unterschiedlichen Rollen mitmachten: Die Schüler/innen jener Klassen, die nur ein Mal ein Tutoring genossen, waren ausschließlich in der Rolle der Tutees und werden daher als *Tutees* bezeichnet. Jene Schüler/innen, die ein Mal oder auch mehrmals ein Tutoring leiteten, werden ab nun als *Tutoren* bezeichnet. Den Ablauf einer solchen, einfachen, CAPT-Sequenz zeigt Abbildung 4.2. Klasse A besteht hier aus ausschließlich aus Tutoren, ist also eine Tutorenklasse, Klasse B hingegen ist die Tuteeklasse.

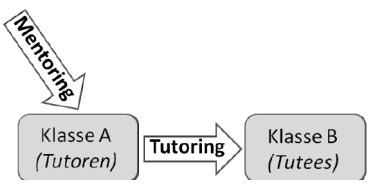


Abbildung 4.2: Ablauf einer einfachen Mentoring-Tutoring-Sequenz

Daneben gibt es aber auch noch Klassen (wie in Abbildung 4.3 die Klasse D), deren Schüler/innen einmal ein Tutoring erlebten, danach ein (eigenes) Mentoring erhielten und in weiterer Folge ein Tutoring mit einer dritten Klasse (E) absolvierten. Die Schüler/innen der Klasse D, die somit in der Doppelrolle waren, werden als *Tutees/Tutoren* bezeichnet. Abbildung 4.3 stellt einen derartigen Ablauf dar.

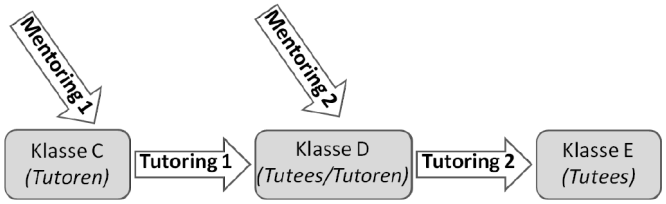


Abbildung 4.3: Mentoring-Tutoring-Sequenz mit Klasse D in Doppelrolle.

Eine dritte Möglichkeit, wie eine CAPT-Intervention aussehen konnte, zeigt Abbildung 4.4. Hier führt eine Klasse (A) mit zwei weiteren Klassen (B, B‘) je ein Tutoring durch.

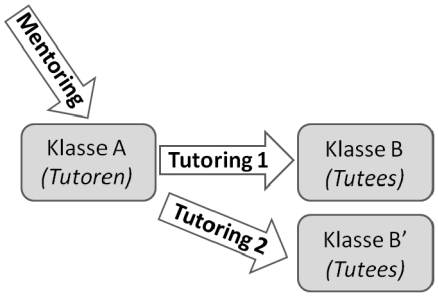


Abbildung 4.4: Tutoring-Sequenz, bei der Klasse A zwei Tutorings durchführt.

Für die Analysen, die in den Kapiteln 7.1 und 7.2 dargestellt sind, wurde der Einfluss der Rollen analysiert. Nun kann man einwenden, dass der Einfluss der Rollen nicht vergleichbar ist, wenn es Tutoren gibt, die ein oder zwei Tutorings durchgeführt haben. Noch problematischer wird es bei jenen Schüler/innen, die die Doppelrolle inne hatten: sie hatten viel mehr Interventionszeit, da sie ein Tutoring, ein Mentoring und ein weiteres Tutoring absolvierten. Es ist also genau zu analysieren, ob mögliche Effekte auf die Interventionszeit, die *time-on-task,* zurückzuführen sind.

4.5. Umgang mit fehlenden Daten

In den ausgewerteten Fragebögen und Wissenstests fehlten immer wieder einzelne Antworten oder ganze Fragebögen. Bevor man statistische Auswertungen beginnt, ist es sinnvoll, sich Gedanken darüber zu machen, wie man damit umgehen möchte. In einem ersten Schritt sind die Gründe für das Fehlen von Werten zu recherchieren, um danach die geeigneten statistischen Mittel anzuwenden.

Die Prozesse, die zu fehlenden Werten (*missing data*) führen, werden üblicher Weise in drei Kategorien geteilt: MCAR, MAR und MNAR (Graham, 2009).

MCAR steht für *missing completely at random*. Der Grund für das Fehlen ist in einem zufälligen Prozess zu finden. Die Gruppe der Versuchspersonen, bei denen Werte fehlen, kann als Zufallsstichprobe angesehen werde. Das ist für alle weiteren Analysen der günstigste Fall, der die geringste Verzerrung oder ungenaue Schätzungen einzelner Parameter nach sich zieht.

MAR steht für *missing at random*. Der Prozess, der zum Fehlen führt ist zwar nicht zufällig, aber bekannt und kann gemessen werden, da es einen Zusammenhang zu anderen Variablen gibt. Für viele weitere Analysen stellt das ebenfalls eine gute Ausgangslage dar und führt zu unverzerrten Parameterschätzungen.

Nicht ganz so gut sieht es aus, wenn es sich um MNAR, *missing not at random*, handelt. Hierbei ist der Grund für das Fehlen von Werten nicht zufällig und liegt in Ereignissen, die nicht gemessen wurden. In diesem Fall ist mit verzerrten statistischen Schätzwerten zu rechnen.

Es ist von Vorteil, wenn die Prozesse, die zu *missing data* führen zugänglich sind, was auch meistens zumindest teilweise der Fall ist (Wayman, 2003).

Ist man sich über die Ursache für *missing data* in einem Datensatz klar, gibt es unterschiedliche Möglichkeiten der Behandlung (Allison, 2009). Eine ältere Methode wäre z.B. das Ersetzen fehlender Werte durch Mittelwerte. Sie ist aber abzulehnen, da in der Folge die Standardabweichungen zu gering geschätzt werden. Konventionelle Methoden sind die *listwise deletion* und die *pairwise deletion*. Die Methode der *listwise deletion* behält nur Datenzeilen, für die alle Werte vorliegen. Die Voraussetzung für eine sinnvolle Anwendung ist, dass die fehlenden Werte MCAR sind. Dann ist das reduzierte

Sample, das nunmehr komplette Fälle enthält, äquivalent zu einem Zufallssample und die Methode führt zu unverzerrten Schätzwerten. Leider wird dabei aber viel von der vorhandenen Information nicht genutzt. *Listwise deletion* kann zu einem erheblichen Datenverlust und damit zum Verlust statistischer Aussagekraft (Power) führen, wenn die Auslassungen ungünstig verteilt sind. Die Methode der *pairwise deletion* schließt jene Variable, für die Daten fehlen aus, benutzt aber alle anderen Werte für den betreffenden Fall. Sie funktioniert gut, wenn fehlende Werte ebenfalls MCAR sind, für MAR führt sie zu verzerrten Schätzern. Dafür werden mehr Daten genutzt als bei der *listwise deletion*.

Eine neuere, attraktive Methode ist die *Multiple Imputation*. Sie liefert unverzerrte Schätzer, ist aber nur für fehlende Werte, die zumindest MAR sind, geeignet. Fehlenden Einträgen werden Werte zugeordnet, die auf Basis der bestehenden Werte, durch eine geschätzte Wahrscheinlichkeitsfunktion für jede Versuchsperson berechnet werden. Dabei geht es nicht darum individuelle Werte bestmöglich zu schätzen oder gar individuelle Vorhersagen zu treffen, sondern darum die Samplecharakteristik zu erhalten (Graham, 2009). Das Programm SPSS, mit dem hier gearbeitet wurde, arbeitet mit einem iterativen EM-Algotithmus (*expectation maximization*), der *Maximum Likelihood* Schätzer berechnet. Damit es sich hierbei nicht um Daten*erfindung* handelt, wird dieser Prozess mehrere Male durchgeführt, und man erhält so einen Vielzahl von imputierten Datensets, über die man abschließend mittelt.

Welche dieser Methoden man anwendet, hängt im Wesentlichen vor allem vom Anteil der *missing data* ab. Fehlen sehr viele Werte, so ist auch die *Multiple Imputation* mit Vorsicht zu genießen, Rost (2007) spricht hier sogar von *multipler Datenerfindung*. Als gerade noch tolerierbare Grenze gibt er einen Anteil von 40 % an fehlenden Werten an. Das Ziel sollte sein, die Anzahl fehlender Werte zu minimieren. Denn sorgsame Planung und Datensammlung kann durch keine statistische Methode ersetzt werden. Fehlen wenige Werte (Richtwert: höchstens 10 %) *„ist es, praktisch gesehen, eigentlich Jacke wie Hose, wie wir mit den missings umgehen“* (Rost, 2007, S. 177). Mit der Methode der *listwise deletion* ist man, seiner Meinung nach, jedenfalls auf sicherem Terrain. Fehlen viele Werte (Richtwert: mehr als 35 %), so sind alle Methoden problematisch. Dazwischen liegt ein Graubereich.

Neben diesen praktischen Richtwerten finden sich in der Literatur auch genauere Angaben: *Listwise deletion* sollte nur bis zu einer Ausschlussquote von 5 % angewandt werden (Lüdtke, Robitzsch, Trautwein, & Köller, 2007), da sonst erhebliche Verzerrungen nicht mehr auszuschließen sind. Außerdem leidet die statistische Power an einer so großen Reduktion der Stichprobe.

Graham (2009) wiederum spricht davon, dass auch 50 % an fehlenden Werten keine Probleme machen, wenn man eine genügende Anzahl von Imputationen berechnet. Auch empfiehlt er persönlich, im Gegensatz zu Rost, bereits ab 5 % fehlender Werte mit *Multiple Imputation* zu arbeiten.

Auf Basis dieses kurzen Abrisses über den Stand der Forschung zum Umgang mit *missing data* wurde die weitere Vorgangsweise für diese Arbeit festgelegt. Zuerst wurden Überlegungen angestellt, warum Daten fehlen. Zwei Punkte waren bei der vorliegenden Studie in diesem Zusammenhang essentiell: Erstens wurde zu Fragebögen, die zu einem der Erhebungszeitpunkte komplett fehlten, recherchiert, dass die Schüler/innen aus Krankheit oder wegen eines schulischen oder außerschulischen Termins verhindert waren. Niemand fehlte aus anderen Gründen, z.B. um die Teilnahme an CAPT und den Befragungen zu vermeiden. Diese Auslassungen können daher als MCAR angesehen werden. Zweitens wurden in einigen Fällen kurze Exitinterviews zu den Wissenstests geführt, mit dem Ziel, die Gründe für einzelne Auslassungen in den Fragebögen zu erfahren. In den meisten Fällen wurde angegeben, dass die Antwort *nicht gewusst* wurde. Somit wurden fehlende Werte in diesen Fragebögen als „falsche Antwort“ gewertet. Diese *missings* können somit als MAR interpretiert werden.

Es wurde entschieden im Rahmen der Prae-Post – Analysen mit *listwise deletion* zu arbeiten, da es im Sinne der obigen Kategorisierung von Rost wenige fehlende Fragebögen zu den einzelnen Testzeitpunkten gab (vgl. Tabelle 4.2). Innerhalb der Fragebögen gab es lediglich vereinzelt fehlende Angaben. Eine Ausnahme davon bildet allerdings der Posttest zum Thema Schatten. Hier hätte bereits mit *Multiple Imputation* gearbeitet werden können. Aus Gründen der Vergleichbarkeit der Daten wurde jedoch für alle Prae-Post-Vergleiche dieselbe Methode angewandt, auch wenn der Preis dafür eine geringere statistische Power für die Vergleiche zum Thema Schatten ist.

Im Rahmen der Analysen zu den Follow-up - Tests wurde hingegen mit der Methode der *Multiple Imputation* gearbeitet, da im Falle einer *listwise deletion* ein nicht unerheblicher Anteil an Fragebögen, vor allem zum Thema Spiegel, nicht zur Analyse herangezogen hätte werden können (vgl. Tabelle 4.2). Auch im Bereich der Elektrizitätslehre und des Schattens fehlen bei den Follow-up Tests erhebliche Anteile an Daten. Der Grund dafür liegt vor allem daran, dass der Follow-up Test gegen Ende des Schuljahres stattfand, wo es für einzelne Schüler/innen immer wieder zur Kollision mit anderen schulischen Terminen kam.

Thema	fehlend beim Praetest	fehlend beim Postttest	fehlend beim Follow-up Test
Elektrizitätslehre	0 %	4,7 %	11,6 %
Spiegel	1,9 %	9,6 %	36,5 %
Schatten	0 %	15,7 %	15,7 %

Tabelle 4.2: Anteile der fehlende Fragebögen zu den einzelnen Testzeitpunkten in Prozenten

Den Ergebnissen vorgreifend sei bereits an dieser Stelle gesagt, dass sich aus dieser Vorgangsweise gut interpretierbare Ergebnisse der Prae-Post-Vergleiche zu den Themen der Elektrizitätslehre und zum Spiegel ergeben. Selbst zum Thema Schatten, das 16 % *missings* aufweist, lässt sich ein hochsignifikant besseres Ergebnis im Posttest nachweisen, wenn auch die Effektstärke etwas davon beeinträchtigt scheint.

5. Messinstrumente

Im Rahmen der Durchführung dieser Studie wurde eine Reihe von Messinstrumenten eingesetzt. Das folgende Kapitel soll eine Übersicht über die verwendeten Instrumente bieten, diese beschreiben und die samplespezifischen Auswahlen begründen.

Die Intervention durch das Tutoring betreffend die Elektrizitätslehre (im ersten Studienjahr) als auch die Optik (im zweiten Studienjahr) basierte darauf, einige wesentliche Basiskonzepte (vgl. Kapitel 4.3) zu vermitteln, abhängig vom Alter und daher vom Wissensstand und den kognitiven Fähigkeiten der Tutees. Bei der Vermittlung dieser Basiskonzepte durch die Tutoren war zu erwarten, dass sich sowohl bei ihnen als auch bei ihren Tutees Veränderungen im Wissensstand ergeben würden. Dieser prognostizierte Wissenszuwachs wurde mit Wissenstests zum jeweiligen Fachgebiet in einem Praetest-Posttest-Follow-up Testdesign festgehalten (vgl. Kapitel 4). Dementsprechend orientieren sich beide Wissenstests ganz konkret an diesen Basiskonzepten und versuchen die Veränderungen bei den Schüler/innen bezüglich dieser Konzepte durch die Intervention zu erfassen.

Hinsichtlich motivationaler und anderer nicht-kognitiver Parameter wie z.B. der Selbstwirksamkeitserwartung wurden zum Teil bestehende Messinstrumente aus der Literatur übernommen. Zur Messung der Motivation selbst wurde kein bestehendes Instrument gefunden, das den Anforderungen genügt hätte. Daher wurde versucht, aus dem Intrinsic Motivation Inventory (Deci & Ryan, 2003) Items zu entnehmen und damit Skalen zu bilden, die auf die Intervention, die adressierte Altersstufe und die sprachlichen Voraussetzungen der getesteten Schüler/innen abgestimmt waren.

5.1. Testinstrument zur Elektrizitätslehre

Das hier verwendete Testinstrument zur Elektrizitätslehre wurde im Rahmen eines bereits länger bestehenden Forschungsprogramms am Austrian Educational Competence Centre Physics, dem AECC Physik, entwickelt (Urban-Woldron & Hopf, 2012). Es entspricht genau den gewünschten Anforderungen an ein Instrument, das sich im Rahmen dieser Studie zielgenau einsetzen lässt, weil es sich ebenso an

Basiskonzepten in der Elektrizitätslehre orientiert, wie die Intervention im Rahmen der Studie. Darüber hinaus war es das definierte Ziel der Entwicklungsarbeit dieses Instruments, dass damit auch möglich sein sollte, bekannte Schülervorstellungen abzubilden und damit ein konzeptuelles Verständnis in der Elektrizitätslehre zu erfassen.

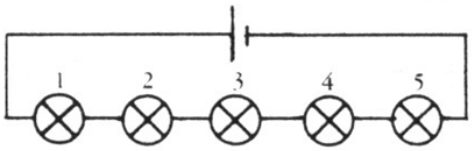
Basis dieses Instrumentes sind verschiedene ältere Tests. Einer der ersten und der wohl bekannteste Test stammt von Christoph von Rhöneck (1986) und wurde zur Erhebung von Vorstellungen über den elektrischen Strom designt. Aus Erfahrungen mit Schüler/innen, die in der Vorbereitungsphase der Studie eingeladen waren, einzelne Items dieses Tests zu lösen und anschließend darüber zu diskutieren, wurde rasch klar, dass dieser Test auch Items beinhaltet, deren Schwierigkeitsgrad weit über das Vermögen selbst interessierter Schüler/innen der Sekundarstufe 1 hinausgeht. Darüber hinaus sind zu diesem Test keine Testkennwerte veröffentlicht. Die ursprüngliche Intention der Studie, im Rahmen der Elektrizitätslehre eine konzeptuelle Entwicklung darstellen zu können, schien mit Rhönecks Test daher nicht möglich zu sein.

Engelhardt und Beichner (2004) entwickelten den *Determining and Interpreting Resistive Electric Circuit Concepts Test* (DIRECT) für Studierende an höheren Schulen und Universitäten. Damit sollen Konzepte und Strategien auf Basis von dichotomen, einstufigen Items (richtig – falsch) erfasst werden. Dieser Test ist allerdings für eine Altersstufe konstruiert, die nicht der untersuchten entsprach.

Aus den genannten Gründen und aus der praktischen Nähe, die die Forschung an einem Testinstrument in der eigenen Arbeitsgruppe bildet, wurde auf den Test zum Erfassen des Verständnisses in der Elektrizitätslehre (Urban-Woldron & Hopf, 2012) zurückgegriffen. Dieser bietet neben einer Diagnosemöglichkeit für Schülervorstellungen durch mehrstufige Items auch den Vorteil, mittlerweile psychometrisch ausreichend validiert zu sein.

Das Verfahren der mehrstufigen Items funktioniert so, dass ein derartiges Item zunächst eine Problemstellung aufwirft. Diese ist in einer ersten Stufe zu beantworten, indem aus verschiedenen Distraktoren gewählt werden kann. Dieses Vorgehen ist zunächst nichts Neues und ist z.B. aus dem Rhöneck-Test (1986) schon bekannt. Als Beispiel sei hier eine Aufgabe (Abbildung 5.1) aus ebendiesem Test vorgestellt:

Fünf gleiche Lämpchen werden in Reihe an eine Batterie angeschlossen:



Kreuzen Sie an, was richtig ist:

1. Lampe 5 leuchtet heller als Lampe 1. ❑
2. Lampe 5 leuchtet so hell wie Lampe 1. ❑
3. Lampe 5 leuchtet schwächer als Lampe 1. ❑

Abbildung 5.1: Aufgabe 4 aus dem Rhöneck-Test (1986)

Urban-Woldron und Hopf entwickelten dieses Testitem insofern weiter, dass die bei ihnen angegebenen Distraktoren der ersten Stufe (a) auf Basis von Interviews mit Schüler/innen gewonnen wurden. Damit wurden die Distraktoren auf der ersten Stufe des ursprünglichen Testinstruments mit Bezugnahme auf bekannte Schülervorstellungen

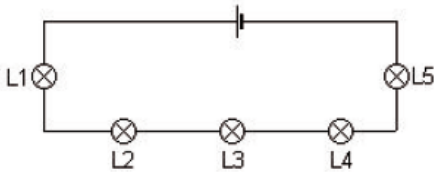
Item 11	**a) Wie hell werden die Glühbirnen leuchten?**	
	L1 leuchtet. Die anderen Glühbirnen leuchten nicht.	
X	Alle Glühbirnen leuchten mit gleicher Helligkeit.	
	L1 und L5 leuchten am stärksten; dann kommen L2 und L4. L3 leuchtet am schwächsten.	
	L3 leuchtet am stärksten; dann kommen L2 und L4. L1 und L5 leuchten am schwächsten.	
	L1 leuchtet am stärksten; dann nimmt die Helligkeit kontinuierlich entlang des Stromkreises ab.	
b)	**Wie erklärst du deine Entscheidung?**	
	Die erste Glühbirne braucht den gesamten Strom; für die anderen ist nichts mehr übrig.	
	Jede Glühbirne verbraucht einen Teil des Stroms, so dass für die nächste weniger übrig ist.	
	Der elektrische Strom wird schwächer je weiter die Glühbirne von der Batterie entfernt ist.	
X	Der elektrische Strom ist an jeder Stelle des Stromkreises gleich.	
	Die Ströme von beiden Polen der Batterie treffen einander bei L3.	

Abbildung 5.2: Weiterentwickeltes, zweistufiges Testitem zum Item aus Abbildung 5.1 nach Urban-Woldron und Hopf (2012, p 210)

ergänzt. Auf der zweiten Ebene ist nun die physikalische Erklärung für die angekreuzte Antwort anzugeben. Damit können neben korrekten Antworten auch „falsch-positive“ und „falsch-negative“ Antworten unterschieden werden: In der klassischen Statistik werden als „falsch-positive“ Antworten solche bezeichnet, bei denen ein Test zwar ein

positives Ergebnis liefert (z.B. Test auf eine Krankheit liefert ein positives Ergebnis), aber bei der Person das Merkmal trotzdem nicht ausgeprägt ist (im Beispiel: Person ist nicht krank, also gesund) (Sachs, 2002). Auf die Situation in diesem Kontext übertragen bedeutet das, dass eine Frage auf der ersten Itemebene zwar richtig beantwortet wurde, aber eine unwissenschaftliche Begründung dahinter steckt, was mit den Distraktoren auf der zweiten Itemebene herausgefunden werden kann. Analog dazu bedeutet „falsch-negativ", dass eine Frage auf der ersten Itemebene zwar falsch beantwortet wurde, aber auf Basis einer korrekten wissenschaftlichen Vorstellung. Mit dieser Unterscheidungsmöglichkeit kann das Testinstrument differenzierter als andere diagnostizieren. Darüber hinaus ist es möglich einer Schülerantwort ein (Fehl-) Konzept zuzuordnen, anstatt bloß festzustellen, dass etwas Falsches angekreuzt wurde.

Den gewählten Antworten in Abbildung 5.2 liegt eine korrekte Antwort *und* eine korrekte Vorstellung zugrunde. Auf der zweiten Itemebene würden Kreuze bei den ersten zwei Möglichkeiten bedeuten, dass der/die Proband/in eine Stromverbrauchsvorstellung hat (von Rhöneck, 1986), bei der dritten Möglichkeit, dass sequenzielles Denken vorliegt (ebd.) und bei der letzten Möglichkeit, dass Strom als Substanz gesehen wird, die von beiden Polen der Batterie kommend sich im Lämpchen trifft (Wiesner, 2004a).

Bezüglich der CAPT-Studie war es nicht nötig, alle Items aus dem *Testinstrument zum Verständnis in der Elektrizitätslehre* zu verwenden. Laut Autoren des Testinstruments (Urban-Woldron & Hopf, 2012) bilden nämlich einzelne Items unterschiedliche Schülervorstellungen ab, was faktorenanalystisch auch untermauert wurde. Entsprechend der in den Interventionen adressierten Konzepte, konnten somit jene Itemblöcke ausgewählt werden, die mit den in der Intervention behandelten Schülervorstellungen übereinstimmten. Damit wurde das Wissen der Schüler/innen vor und nach der Intervention abgefragt. Aus dem ursprünglichen Testinstrument wurden die Items 1 bis 10 verwendet. Davon wurden aber lediglich die Items 1 bis 5 zur Auswertung herangezogen. Der Grund dafür ist, dass die Items 6 bis 10 keinen Konzepten entsprachen, die in der CAPT Intervention Thema waren.

Da hier nicht der komplette Test eingesetzt wurde, wurden die Reliabilitäten des Teiltests überprüft. Es ergaben sich Werte von α = 0,78 für den Praetest (Item 1 bis 5 bei summativer Zählweise) und α = 0,85 für den Posttest (Item 1 bis 5).

Die Items 6 bis 10 korrespondierten nicht mit den tatsächlich behandelten Themen der CAPT Intervention. Sie wurden jedoch verwendet, da die Intervention in gewisser Hinsicht nach oben offen war und es Schüler/innen frei stand auch eigene Ideen einzubringen. In diesem Falle sollten genügend Testitems vorhanden sein. Gleichzeitig sind die Items, von denen erst a posteriori fest stand, dass sie nicht gebraucht wurden, eine gute Möglichkeit, den Test einer Validitätsprüfung zu unterziehen. Es ist zu sehen, dass sich für die Items 6 bis 10, die letztlich nicht Thema der CAPT Intervention waren, die Anzahl der korrekt beantworteten Fragen in den Praetests wenig von der der Posttests unterscheidet (vgl. Abbildung 5.3), bzw. nicht notwendiger Weise eine Steigerung (z.B. Item 7) zu erkennen ist. Im Gegensatz dazu liegt die Anzahl der korrekt beantworteten Items in den Posttests über jener der Praetests für alle Items, die sich auf Themen der Intervention beziehen. Mit anderen Worten: Abbildung 5.3 liefert gute Hinweise darauf, dass die Themen der Intervention und die Testitems gut aufeinander abgestimmt sind und der Test zumindest für die untersuchte Stichprobe valide ist.

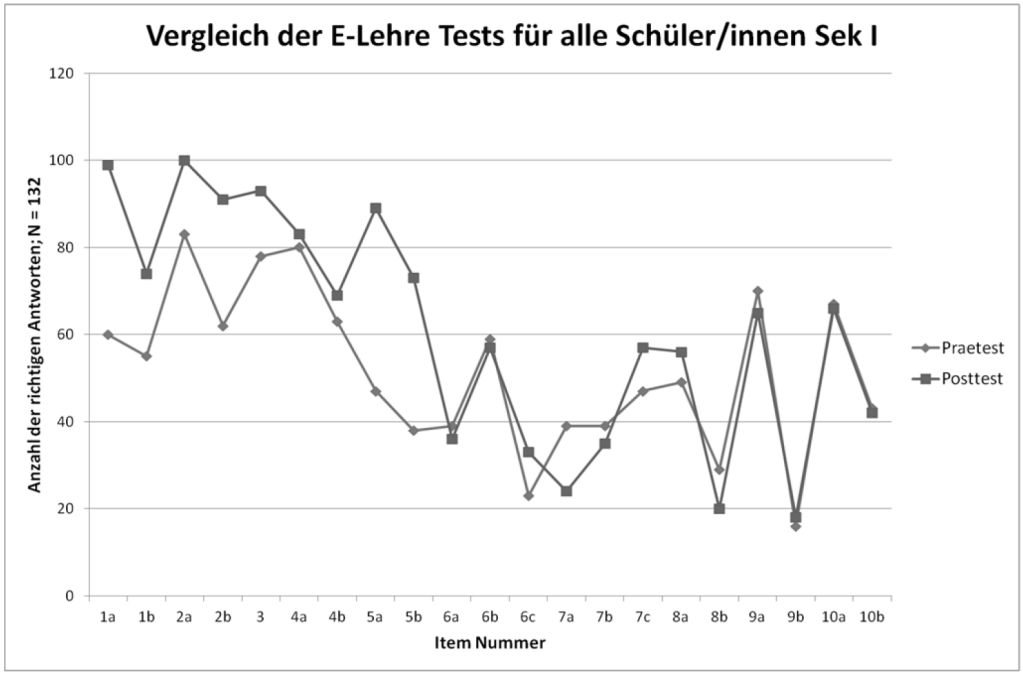


Abbildung 5.3: Anzahl der Schüler/innen, die im Prae- und Posttest das entsprechende Teilitem korrekt beantworteten. Items 1a bis 5b entsprechen Themen der Intervention, Items 6a bis 10b nicht. Die Verbindungslinien dienen lediglich der besseren Orientierung.

Die nahezu identischen Ergebnisse der Prae- und Posttests bei den Items 8 bis 10, im Gegensatz zu denen der Items 1 bis 5, sind ein Indiz für die Reliabilität des Testinstruments und können gleichzeitig als erster Hinweis auf die Lernwirksamkeit der CAPT Intervention gesehen werden, ohne genauen Analysen aus Kapitel 7.1 vorgreifen zu wollen.

Bei näherer Betrachtung fällt auf, dass bei diesem ersten Überblick über die Testergebnisse bei den einzelnen Testitems die oben beschriebene Zweistufigkeit nicht berücksichtigt wurde. In der Tat wurde dieses Merkmal des Testinstruments zur Elektrizitätslehre auch in den folgenden Auswertungen nicht berücksichtigt. Statt der zweistufigen, konzeptorientierten Auswertung wurden Summenscores über die Teilitems gebildet. Der Unterschied besteht darin, dass bei der konzeptuellen Auswertung ein Punkt vergeben wird, wenn auf beiden Itemebenen (a und b) korrekt geantwortet wurde. Bei einem Summenscore wird für die Teilitems jeweils ein Punkt vergeben, unabhängig davon ob die zweite Itemebene korrekt oder inkorrekt beantwortet wurde (*partial credits*).

Nach langen Diskussionen und zahlreichen Versuchen in die eine oder andere Richtung hat sich herausgestellt, dass die Vorgangsweise der Auswertung der mehrstufigen Items, wie von den Autoren vorgeschlagen, hier nicht zielführend ist. Das Testinstrument vermag zwar mehr zu leisten, als wofür es tatsächlich eingesetzt wurde: Mit diesem Testinstrument sollte nämlich auch ein Konzeptwechsel nachgewiesen werden können, und nicht bloß eine unspezifische erhöhte *task completion rate* dargestellt werden. Die Datenlage in Verbindung mit der Intervention, so wie sie stattgefunden hat, legt jedoch hier ein anderes Vorgehen nahe. Es wurde eben nicht der gesamte Unterricht zur Elektrizitätslehre evaluiert, sondern der Blick auf wenige bestimmte Konzepte gerichtet. Das liegt an der eher kurzen Dauer er Intervention und daran, dass, angepasst an die Altersstufen der Tutees, teilweise unterschiedliche Themen behandelt wurden, die allerdings im Kernbereich Überschneidungen hatten. Nun ist es in diesem Fall aber aus messtheoretischer Sicht nicht sinnvoll, einen Teil eines Testinstruments einzusetzen, der nur ein oder zwei Konzepte erfasst und daraus auf einen eingetretenen oder eben nicht eingetretenen konzeptuellen Fortschritt zu schließen, um auf Basis dessen die Güte der Intervention zu beurteilen. Das Raster erscheint hier zu grob, man wird rasch an Boden-

oder Deckeneffekte stoßen. Um das zu vermeiden ist es sinnvoller, innerhalb dieser Konzepte mittels eines Summenscores die jeweilige Ausprägung zu quantifizieren und auf Basis dessen zu beurteilen, ob sich durch die Intervention eine Weiterentwicklung *innerhalb* des Konzeptes ergibt oder nicht. Die hier angewandte, summative Zählart der Punkte soll demnach abbilden, *wie viel* von einem Konzept erfasst wurde: Ergeben sich mehr Punkte, wird ein Test dahingehend interpretiert, dass vom jeweiligen Konzept mehr erfasst wurde oder dass das Konzept als Erklärungsbasis belastbarer ist als bei einem geringeren Punktescore. Es wird daher auch in fachlich kritischen Phasen, bei schwierigen Fragestellungen, das Alltagskonzept weniger häufig zu Rate gezogen werden. Somit steht auch bei Verwendung von Summenscores einer differenzierteren Interpretation der Weg offen, und es kann eine Entwicklung in eine bestimmte Richtung herausgelesen werden.

5.2. Testinstrument zur Optik

Die hier berichtete Studie war über zwei Schuljahre, von 2010 bis 2012, angelegt. Thema im zweiten Schuljahr war die Optik. Die Ausgangslage betreffend ein Testinstrument zur Optik war eine gänzlich andere als in der Elektrizitätslehre. Daran war die weitere Vorgehensweise anzupassen.

Zwar gibt es in der Optik, ähnlich wie in der Elektrizitätslehre, eine gute Basis an bereits erforschten und bekannten Schülervorstellungen. Das ist ein großer Vorteil und nicht auf allen Gebieten der Physik selbstverständlich. Man denke nur an z.B. ein unmittelbar „benachbartes“ Gebiet, die nicht-sichtbare Strahlung, zu der, abgesehen von einigen neueren Publikationen (z.B. Neumann & Hopf, 2012), international nur wenig Forschungsergebnisse vorliegen.

Die CAPT-Intervention im Rahmen des zweiten Studienjahres sollte, abgesehen vom Thema, ähnlich verlaufen: Es waren dieselben Schulen mit fast gleichen Klassen beteiligt, was eine bestimmte Altersgruppe an Proband/innen und denselben sozio-ökonomischen Rahmen bedeutete. Darüber hinaus sollte sich auch diese Intervention an Basiskonzepten, diesmal zum Themenbereich der Optik orientieren (vgl. Kap.4). Die

CAPT Intervention beinhaltete thematisch vor allem die Bereiche Schatten und Spiegel und die damit verknüpften Schülervorstellungen.

Die große Herausforderung besteht darin, dass es meines Wissens bis dato kein psychometrisch valides und reliables Fragebogeninstrument zur Optik gibt, das sich an Schülervorstellungen orientiert. Zwar gibt es am AECC Physik Forschung zur Konstruktion eines solchen Testinstruments, das Instrument selbst ist aber noch in der Entwicklungs- und Erprobungsphase. Aus diesem Grund konnte zum Zeitpunkt der Durchführung der Studie lediglich auf einige Items rund um den Themenbereich „Spiegel“ zurückgegriffen werden, die bereits einer ersten Schleife aus empirischer Erprobung und Redesign unterzogen waren (Haagen-Schützenhöfer & Hopf, 2011; Haagen-Schützenhöfer, Rottensteiner, & Hopf, 2012). Während es zum Themenkomplex Spiegel somit einige erprobte Items gab, gibt es bis dato kaum gründlich getestete Items zum Themenkomplex Schatten.

Potenzielle Schwierigkeiten bei Schatten-Items liegen darin, dass einerseits in der hier berichteten Studie Schatten mit eher jüngeren Schüler/innen behandelt wurde. Andererseits sind für die Darstellungen dreidimensionale Abbildungen nötig, die eine große Herausforderung an die räumliche Vorstellungskraft der Schüler/innen darstellen. Aus der bisherigen Arbeit mit Schüler/innen der Sekundarstufe 1 geht hervor, dass gerade das eine große Hürde darstellt (Haagen-Schützenhöfer, et al., 2012). Ein weiteres Problem besteht darin, dass es in der Optik, im Unterschied zur Elektrizitätslehre, keine gängige, zweidimensionale Symbolik gibt, die die Darstellungen vereinfacht. Nun kann man darüber streiten, ob die Symbolik in der Elektrizitätslehre Lernschwierigkeiten erzeugt oder nicht. Faktum ist aber, dass damit lange, die Abbildungen erklärende Texte überflüssig werden, während sie in der Optik vonnöten sind. Das stellt vor allem im Bereich des sinnerfassenden Lesens große Herausforderungen an die Proband/innen dar und erhöht somit die Itemschwierigkeit.

Da nach ersten Auswertungen zur Elektrizitätslehre davon ausgegangen werden konnte, dass die CAPT-Intervention lernwirksam für Tutoren und Tutees ist, wurde in weiterer Folge versucht, zu den beiden Themenkomplexen Schatten und Spiegel Items zu finden, die an Schülervorstellungen der Intervention anknüpfen, von der gewählten Darstellung her möglichst verständlich und gleichzeitig wenig textlastig sind.

Der Schwerpunkt der Auswertungen konnte im Bereich der Optik nicht auf einer Analyse der Lernwirksamkeit von CAPT hinsichtlich der Rollen liegen, da die Aufteilung des Samples nach Rollen *und* Themen zu kleine Gruppen erzeugt hätte, als dass statistische Tests aussagekräftige Ergebnisse liefern könnten. Es wurde vielmehr darauf fokussiert, ob sich CAPT als Lernmethode auch für möglichst viele weitere Themen eignet, eben für die beiden berichteten Bereiche der Optik. Entsprechend dem Fehlen eines ausgereiften Testinstruments zur Optik wurde der Fokus der Analysen darauf gelegt, ob es überhaupt zu einem Lernzuwachs in den einzelnen Klassen kommt und ob dieser, wenn er eintritt, mit bestimmten Faktoren verknüpft werden kann.

An einer der Tutorenklassen, die an der Studie beteiligt waren, wurde eine Vorstudie zum Fragebogeninstrument der Optik durchgeführt.

Vorstudie zur Optik an Klasse 5

Die Klasse 5 war die einzige Oberstufenklasse in der Studie und bestand aus lediglich neun Schüler/innen, da es sich um den naturwissenschaftlichen Teil einer typengemischten Klasse handelte. Mit dieser Klasse konnte relativ viel Zeit verbracht werden (vier Unterrichtseinheiten à 50 min). Dabei wurde zum einen das Design des Mentorings erprobt und anhand von Schülerrückmeldungen verbessert.

Zum anderen wurden die Fragebogenitems des Wissenstests Optik in dieser Klasse pilotiert und anhand von Exitinterviews mit den Schüler/innen verbessert. Den Schüler/innen wurden die Testitems als Praetest ausgehändigt. Danach fand ein erweitertes Mentoring statt, das beide Themenbereiche, Spiegel und Schatten, abdeckte. Im Anschluss daran wurden mit den Schüler/innen die Auflösungen der Testfragen besprochen. In einer offenen Interviewform wurden die grafischen Darstellungen detailliert diskutiert, und es wurde erfragt, was die Schüler/innen darin erkennen können. Auch nach einer mehrstündigen Intervention zur Optik korrespondierte das nicht immer mit dem, was sich die Autoren der Items vorgestellt hatten. Des Weiteren wurden die Schüler/innen nach der Textverständlichkeit befragt. Als Expert/innen für das Lernen mit Jüngeren sollten sie auch beurteilen, inwiefern sie die Texte als für Schüler/innen der Sekundarstufe 1 für verständlich erachteten. Gemeinsam wurden Darstellungen abgeändert bzw. unübliche Wörter in Textpassagen durch schülernähere, aber trotzdem nicht umgangssprachliche, ersetzt. Nach der ersten

Verbesserungsphase wurden die nun erstellten Testitems den Schüler/innen der Sekundarstufe 1 als Prae- und Posttests vorgelegt.

Beispielitem zum Thema Schatten

Auf Basis des Artikels von H. Wiesner über Vorstellungen von Grundschülern über Schattenphänomene (2004c) wurde ein Item entworfen, das in Abbildung 5.4 abgebildet ist. Die Fragestellung dazu lautete: „Kreise den richtigen Schatten ein!“.

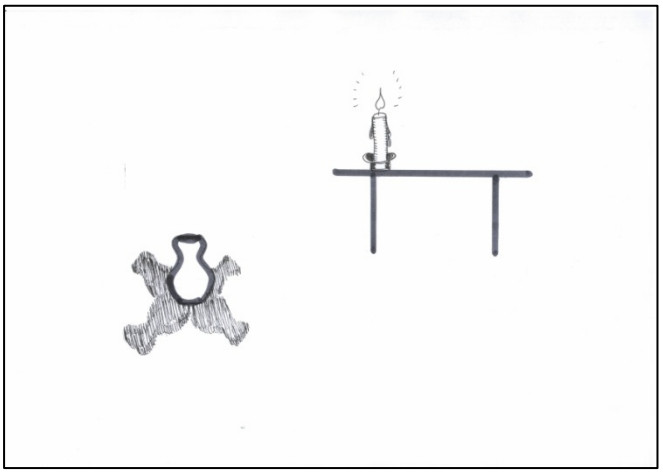


Abbildung 5.4: Beispielitem aus dem Schattentest. Nach (Wiesner, 2004c, S. 71).

Dieses Item enthält die Lichtquelle (Kerze) und vier mögliche Schattenpositionen, bei denen sich die Schüler/innen für eine korrekte entscheiden müssen. Dieses Item wurde als korrekt beantwortet gewertet und dafür ein Punkt vergeben, wenn der richtige Schatten als einziger markiert wurde.

Dieses Item zielt darauf ab, dass Schüler/innen zur korrekten Beantwortung wissen müssen, dass sich Licht geradlinig ausbreitet. Es schließt somit unmittelbar an das Basiskonzept der geradlinigen Ausbreitung von Licht an. Ferner müssen sie wissen, dass Schatten Lichtmangel bedeutet. Schatten sehen wir dort, wo wenig oder kein Licht hinfällt.

Die Bewertung der jeweiligen Schülerantworten wurde innerhalb der Arbeitsgruppe mit zwei weiteren Wissenschaftlern expertenvalidiert. Insgesamt verfügt der Wissenstest zum Schatten über 8 Items, für die, anhängig von der Komplexität der Fragestellung, ein oder zwei Punkte vergeben wurden. Die Items, für die zwei Punkte vergeben wurden, waren solche, wo es neben einer korrekten Antwort auch die Möglichkeit gab, diese

grafisch oder verbal zu begründen. War lediglich eine der beiden Fragestellungen korrekt beantwortet, wurde dafür ein Punkt vergeben. Es wurde daher mit einem *partial credits* Verfahren gearbeitet. Insgesamt konnte man bei diesem Wissenstest maximal 12 Punkte erzielen.

Die post-hoc berechnete Reliabilität dieses Test lag bei $\alpha = 0{,}66$, was als zufriedenstellend bewertet werden kann, da es sich nicht um ein ausgereiftes, in mehreren Zyklen empirisch erprobtes und verbessertes Messinstrument handelt.

Beispielitem zum Thema Spiegel

Der Wissenstest zum Thema Spiegel basierte auf einem bereits in der Erprobungsphase befindlichen Testinstrument (Haagen-Schützenhöfer & Hopf, 2011), aus dem aber ebenfalls, wie im Falle des Test zur Elektrizitätslehre nur einige Items, entsprechend der adressierten Schülervorstellungen, entnommen wurden. Abbildung 5.5 stellt ein derartiges Item dar, das in seiner ursprünglichen Form von H. Wiesner stammt (1992) und in der hier abgebildeten Form eine Überarbeitung von C. Haagen-Schützenhöfer (2011) darstellt. Dieses Item testet das Basiskonzept ab, wo sich der Ort des

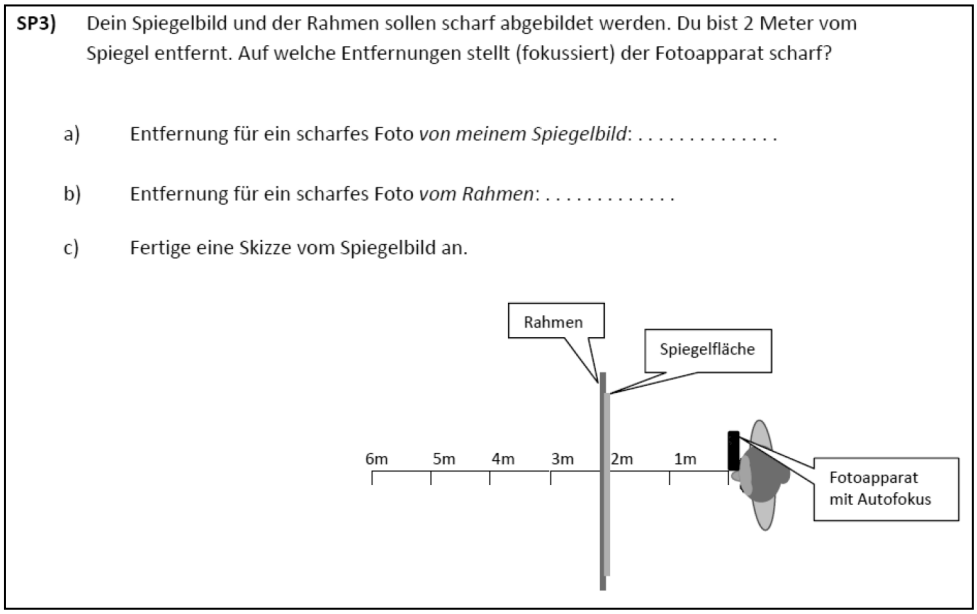


Abbildung 5.5: Beispielitem aus dem Wissenstest zum Thema Spiegel (Haagen-Schützenhöfer, 2011)

Spiegelbildes befindet. Eine differenzierte Diagnose möglicher Fehlkonzepte (z.B. dass das Spiegelbild auf dem Spiegel liegt) sollte damit bereits möglich sein.

Da die Teilitems voneinander relativ unabhängig sind, wurden hier, in Übereinstimmung mit der Autorin, *partial credits* vergeben und diese summiert.

Insgesamt beinhaltet dieser Wissenstest fünf Fragen, innerhalb derer zwei bis drei Unterfragen zu beantworten sind. Das ergibt, die unterschiedliche Komplexität dieser Fragen berücksichtigend, einen maximale Punkteanzahl von 11 Punkten.

Auch zu diesem Test wurde post hoc eine Analyse der Reliabilität durchgeführt, die ein Cronbach's α von 0,63 ergab, was in diesem Zusammenhang wiederum als zufriedenstellend gewertet wurde.

5.3. Messinstrumente zur Erfassung einiger nicht-kognitiver Variablen

Parallel zu den oben beschriebenen Wissenstests zur Elektrizitätslehre bzw. zur Optik, wurde eine Reihe nicht-kognitiver Variabler erhoben. Damit sollen rein deskriptiv Zusammenhänge zwischen Testergebnissen, also zwischen kognitiven, und nicht-kognitiven Parametern erfasst werden, um eine Beschreibung von CAPT in multi-perspektivischer Hinsicht zu ermöglichen.

Tabelle 5.1 gibt einen Überblick über die verwendeten Messinstrumente, die dann in weiterer Folge näher beschrieben werden. Die unterschiedlichen Testzeitpunkte ergeben sich aus der inhaltlichen Ausrichtung des jeweiligen Fragebogens. So wurde z.B. der Fragebogen zum Lernen im Fach konstruiert, um die bereits bestehende Motivation für das Fach Physik zu erfassen. Der Fragebogen zur Selbstwirksamkeitserwartung bezieht sich im Gegensatz dazu auf eine kommende Herausforderung.

Fragebogen	Testzeitpunkt	Autoren
Lernen im Fach	vor dem Mentoring	Müller, Hanfstingl, Andreitz (2007)
Aktuelle Motivation	nach dem Mentoring	Rheinberg, Vollmeyer, Burns (2001)
Selbstwirksamkeits-erwartung	vor dem Tutoring (nur Optik)	Schwarzer, Jerusalem (1999a)
Motivation	nach dem Tutoring	Korner, Urban-Woldron, Hopf (2012)

Tabelle 5.1: Überblick über die eingesetzten Fragebogeninstrumente zur Erfassung nicht-kognitiver Variablen

5.3.1. Fragebogen zum Lernen im Fach

Dieser Fragebogen von F.H. Müller et al. (2007) stellt eine Bearbeitung des *Academic Self-Regulation Questionnaires* (Ryan & Connell, 1989) dar. Dieser wiederum basiert auf der Selbstbestimmungstheorie (SDT) der Motivation (Deci & Ryan, 1985, 1993, 2008). Die SDT beschreibt zwei komplementäre Ausprägungen der Motivation: die extrinsische und die intrinsische. Die extrinsische Motivation selbst wird in vier Stufen unterteilt, die sich im Grad der Ausprägung der Autonomie unterscheiden: beginnend mit der externalen Regulation (geringe Autonomie, Ursächlichkeit außerhalb des Individuums) über die introjizierte Regulation, die identifizierte Regulation bis hin zur integrierten Regulation (große Autonomie, Ursächlichkeit bereits im Individuum selbst). Eine detaillierte Beschreibung dazu findet sich in Kap. 2.3.

Im Fragebogen zum Lernen im Fach(F. H. Müller, et al., 2007, S 5), wie auch im *Self-Regulation Questionnaire*, werden diese vier Stufen der extrinsischen Motivation vier Skalen zugeordnet. Dabei wird auf eine eigene Skala zur intrinsischen Motivation verzichtet, da sie besonders bei jüngeren Schüler/innen faktoranalytisch nicht trennscharf von der integrierten Regulation unterschieden werden kann (Vallerand, 2000). Jede Skala im Fragebogen zum Lernen in Fach enthält vier oder fünf Items, wovon jedes in einer sechsteiligen Likertskala abgefragt wird.

Die Mittelwerte dieser vier Skalen werden dann zum sogenannten Selbstbestimmungs-index (SDI) nach folgender Formel zusammengefasst:

$$SDI = 2 \times (intrinsische\ Regulation) + identifizierte\ Regulation \\ - introjizierte\ Regulation - 2 \times (externale\ Regulation)$$

Ein positives Vorzeichen erhalten die selbstbestimmten Regulationsstile, ein negatives die externalen Regulationsstile. Dabei soll der Faktor 2 jeweils die stärkere Ausprägung der (vorhandenen oder eben fehlenden) Autonomie darstellen. Der SDI liegt somit im Intervall $[-15; +15]$.

Die Autoren beschreiben, dass konfirmatorische Faktoranalysen eine gute Trennung der einzelnen Skalen ergeben. Zum Validierungsprozess geben die Autoren an, dass als Außenkriterium die drei selbstbezogenen Variablen Fachinteresse, fachliche Sorge und fachliches Selbstkonzept aus der PISA-2003-Studie (Haider & Reiter, 2004) übernommen

wurden. Dass die errechneten Korrelationen die erwarteten Zusammenhänge zeigen, wird als ein erster Hinweis auf die Konstruktvalidität gewertet. Allerdings lassen die Autoren die Frage offen, inwiefern extrinsische oder intrinsische Motivation mit den drei angeführten Variablen zusammenhängen.

Die Reliabilitäten des Fragebogens (Cronbach α) liegen für die einzelnen Subskalen zwischen $\alpha = 0{,}75$ und $\alpha = 0{,}92$ und werden als zufriedenstellend bewertet. Bei Aggregation auf Klassenebene liegen die Werte sogar noch darüber. Der Einsatz des Fragebogens wird aufgrund von Schüler- und Lehrerrückmeldungen ab etwa 11 Jahren empfohlen. Für Grundschüler/innen wird vom Einsatz abgeraten, da die Formulierungen in den einzelnen Items ein zu hohes Abstraktionsniveau voraussetzen. Diese Empfehlung war einer der Gründe, warum sich die Forschungsfragen dieser Arbeit auf Untersuchungen der Sekundarstufe 1 beschränkten.

In der hier beschriebenen Studie wurde der Fragebogen fast unverändert übernommen. Lediglich zwei Dinge wurden abgeändert: der Begriff „Fach“ wurde durch „Physik“ ersetzt und, da dieser Fragebogen quasi den Auftakt der Befragungen in den Klassen darstellte, wurden zu Beginn demografische Variable abgefragt: das Geschlecht, die letzte Physiknote in einem Zeugnis und die Muttersprache im Sinne der daheim gesprochenen Sprache.

5.3.2. Fragebogen zur aktuellen Motivation

Der hier eingesetzte Fragebogen der Autoren Rheinberg, Vollmeyer und Burns (2001) nennt sich im Volltitel *FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen (Langfassung 2001)*. Er bezieht sich, anders als der *Fragebogen zum Lernen im Fach*, auf das Modell der Leistungsmotivation (Risiko-Wahl-Modell) von Atkinson.

Der Begriff der *aktuellen Motivation* soll den personenspezifischen Faktor des Motivs innerhalb einer gegebenen Situation erfassen. Die klassische Motivationspsychologie geht davon aus, dass sich Motivation aus der Interaktion zwischen personenspezifischen und situationsspezifischen Merkmalen ergibt (Rheinberg, et al., 2001). Die ersteren werden als *Motive* bezeichnet und beinhalten die Klasse von Anreizen, die eine Person bevorzugt. Um aktuelle Motivation zu generieren, die dann auch ein bestimmtes

Verhalten auslöst, müssen diese Motive zur Situation passen. Mit dem FAM wollen die Autoren weg von personenspezifischen Motiven gehen und in unterschiedlichsten Lernsituationen die aktuelle Motivation erfassen.

Im FAM werden die einzelnen 7-stufig likert-skalierten Items zu vier Subskalen zusammengefasst: Misserfolgsbefürchtung, Erfolgswahrscheinlichkeit aus Sicht des Probanden, Interesse an der Situation und die empfundene Herausforderung, die die spezielle Situation mit sich bringt. Der Faktor *Misserfolgsbefürchtung* soll den negativen Anreiz einer Situation, der durch einen möglichen Druck entsteht, abbilden. Der Faktor *Erfolgswahrscheinlichkeit* bildet die eigenen Einschätzung des Probanden ab, wie sicher er/sie ist, hier gut anzuschneiden, weil die eigenen Fähigkeiten hoch und/oder die Aufgabe als leicht eingeschätzt wird. Diese Skala bezieht sich also auf ein Konstrukt, das dem der Selbstwirksamkeitserwartung sehr ähnlich ist. Der Faktor *Interesse* betrifft das Interesse im Sinne der Interessenstheorie von Krapp und Ryan (Deci & Ryan, 2002; Krapp, 2002; 2002) an den spezifischen Inhalten einer Aufgabe, während der Faktor *Herausforderung* erfasst, wie passend der Schwierigkeitsgrad einer Aufgabe eingeschätzt wird, d.h. ob eine Aufgabe weder als zu leicht noch als zu schwer empfunden wird.

Diese vier Faktoren lassen sich in faktoranalytischen Untersuchungen zufriedenstellend voneinander trennen. Die Autoren des FAM (Rheinberg, et al., 2001) berichten, dass die Reliabilitäten (Cronbach α) der einzelnen Skalen, je nach Stichprobe und daher getesteter Aufgabensituation, zwischen $\alpha = 0{,}66$ und $\alpha = 0{,}90$ liegen. Sie geben an, dass sich Unterschiede in den Aufgaben mit dem FAM abbilden lassen, wenn sich diese z.B. durch das Niveau der Problemlösungsprozesse unterscheiden. Bezüglich der Validität des Instrumentes geben die Autoren an, dass sich neben bekannten kognitiven Faktoren bereits Einflüsse von FAM-Faktoren auf Lernergebnisse nachweisen lassen haben.

5.3.3. Fragebogen zur allgemeinen Selbstwirksamkeitserwartung

Der Fragebogen zur allgemeinen Selbstwirksamkeitserwartung (SWE) wurde bereits 1981 von Schwarzer entwickelt und umfasste in der ursprünglichen Version 20 Items, die später auf 10 Items reduziert wurden (Schwarzer, 1993). Jedes Item wird in einer vierteiligen Likertskala abgefragt.

Dieses Fragebogeninstrument wurde in der Folge in 30 Sprachen übersetzt und steht nun online zur Verfügung (Schwarzer & Jerusalem, 1999a). Das Konstrukt, das diesem Instrument zugrunde liegt, ist Banduras *perceived self-efficacy*, das Selbstwirksamkeitskonzept (Bandura, 1977, 1997). Das Instrument erfasst die meist stabile Erwartungshaltung einer Testperson, eine schwierige Situation meistern zu können auf Basis der Überzeugung, dass man den Erfolg der eigenen Kompetenz zuschreiben kann und nicht extrinsische Faktoren über Gelingen oder Misslingen entscheiden (vgl. Kapitel 2.3).

Faktoranalytisch zeigen die zehn Items eine sehr stabile Eindimensionalität. Die Autoren geben an, dass die Reliabilitäten für internationale Stichproben zwischen 0,76 und 0,90 liegen. Da die Ergebnisse likert-skalierter Items oft vom kulturellen Kontext abhängig sind, wurde sogar für den chinesischen Kulturkreis eine adaptierte Version des Instruments erstellt und getestet. Für rein deutsche Stichproben lagen die Reliabilitäten zwischen 0,80 und 0,90 (Schwarzer & Jerusalem, 1999b; Schwarzer, Mueller, & Greenglass, 1999).

Für die Zwecke der hier vorgestellten Studie wurde das Instrument lediglich im zweiten Studienjahr eingesetzt, um Hinweise für die Validität der anderen verwendeten Instrumente zu bekommen, insbesondere für das Motivationstestinstrument, über das in Kapitel 6 ausführlich berichtet wird.

6. Konstruktion eines Fragebogens zur Motivation

Im Rahmen der hier vorgestellten Studie soll im Sinne einer umfassenden Beschreibung der Methode CAPT und ihrer Lernwirksamkeit neben kognitiven Parametern die Lernmotivation der Schüler/innen beschrieben werden. Daher wird in diesem Kapitel ausführlich berichtet, aus welchen Gründen bestehende Instrumente als nicht passend empfunden wurden, wie die Entwicklung eines eigenen Instrumentes von statten ging und mit welchen Schwierigkeiten sie verbunden war.

6.1. Anforderungen an ein Messinstrument zur Motivation

Eine der Fragen, die im Rahmen dieser Studie verfolgt werden sollte, war, ob die Unterrichtsmethode CAPT, in Abhängigkeit von der Rolle, die die Schüler/innen im Tutoringprozess innehatten, Auswirkungen auf einen möglichen Wissenszuwachs der Schüler/innen hat. Um die Ursachen dieses potenziellen Wissenszuwachses genauer lokalisieren zu können, war es geplant, auch nicht-kognitive Parameter zu erfassen. Einer dieser nicht-kognitiven Parameter soll, wie in Forschungsfrage 7 ausgeführt, erfassen, wie lern-motivierend CAPT als Methode wahrgenommen wird. Ebenso ist es hinsichtlich eines *Conceptual Change*, der auch multi-perspektivisch gesehen werden kann, und hinsichtlich einer Implementierung von CAPT in den Regelunterricht wünschenswert, die Motivation zu erfassen.

Um das beurteilen zu können, bedarf es einer akkuraten und differenzierten Diagnose der Motivation der Schüler/innen. Eben dafür ist ein Messinstrument von Nöten.

Steht ein Messinstrument zur Motivation zur Verfügung, besteht außerdem die Möglichkeit, ein Abschneiden der Schüler/innen im Wissenstest mit der Motivation in Verbindung zu setzen, die sie für diese Tätigkeit aufbrachten. Eine mögliche Forschungshypothese könnte hier lauten, dass höher motivierte Schüler/innen besser abschneiden oder dass sie ihre Tutees besser betreuen und diese daher besser abschneiden. Darüber hinaus ist es interessant zu erfahren, welche Aspekte der Motivation durch CAPT angesprochen werden, um somit herauszufinden, *was* an dieser Methode für die Schüler/innen motivierend ist und wie sich das genau auswirkt.

Ein Messinstrument, das im Zusammenhang mit dieser Studie verwendet wird, soll daher eine differenzierte Sicht auf die Motivation der Schüler/innen ermöglichen. Aus diesem Grunde wurde als zugrunde liegendes Konstrukt für die Motivation unter den vielen Motivationstheorien, die unterschiedlichste Aspekte abbilden, die SDT gewählt (siehe Kapitel 2.3). Sie beschreibt eben diese geforderte differenzierte Sicht auf die Motivation der Schüler/innen, hinsichtlich der Ursachen ebenso wie hinsichtlich der emotionalen Qualität.

Die SDT unterscheidet zwischen intrinsischer Motivation und insgesamt vier Formen extrinsischer Motivation, und somit auch zwischen den Loci, aus denen die Motivation kommt. Bei intrinsisch motiviertem Handeln ist der Lokus der Entscheidung, diese Handlung zu beginnen im Selbst der handelnden Person verortet. Bei extrinsisch motiviertem Handeln kann er auch außerhalb der Person liegen. Diese Differenzierungen erscheinen gerade im Kontext eines Peertutorings wesentlich zu sein.

Die SDT scheint vom Konstrukt her gewissermaßen geeignet zu sein, einen Brückenschlag zwischen quantitativer und qualitativer Forschung zu bilden. Einer der Kritikpunkte an quantitativer Forschung ist der, dass sie die Proband/innen auf einige wenige Variable reduziert (Lamnek, 2005). Diese Variablen werden zwar theoriebasiert, bleiben aber dennoch willkürlich. Die Auswahl als solche schränkt die Komplexität des Forschungsgegenstandes bereits ein und verkürzt ihn. Ein ganzheitliches Bild ist somit nicht zu erwarten. Die SDT hingegen stellt bereits als Theorie eine derart umfassende und individualisierende Betrachtung zur Verfügung, dass selbst ein quantitatives Messinstrument, das auf ihr basiert, sehr individuelle Aussagen nach den Ursachen zulässt.

Eine psychometrische Umsetzung des Motivationskonstrukts der SDT liefert das Intrinsic Motivation Inventory, das IMI (Deci & Ryan, 2003), ein Fragebogen in englischer Sprache.

Die in dieser Studie untersuchte Stichprobe bestand zum Großteil aus Schüler/innen der Sekundarstufe 1 (Alter 10 bis 14 Jahre). Daher sollte ein geeignetes Messinstrument auf die sprachlichen und kognitiven Fähigkeiten dieses Alters abgestimmt sein. Von diesen Schüler/innen besuchte die Mehrheit eine Hauptschule (vgl. Kapitel 7.1.1), was im untersuchten städtischen Ballungsraum in Österreich zu einem hohen Anteil an

Risikoschüler/innen, speziell im Bereich des Lesens (Schmirch, 2009; Statistik-Austria, 2014) führt (vgl. auch S. 56). Ein erheblicher Teil dieser Jugendlichen wies darüber hinaus einen Migrationshintergrund auf, entsprechend der Definition der OECD (Breit, 2009): Demzufolge gelten als *Migrant/innen erster Generation* Schüler/innen, die, ebenso wie ihre beiden Elternteile, im Ausland geboren wurden. *Migrant/innen zweiter Generation* sind Schüler/innen, die zwar selbst im Land geboren wurden, aber beide Elternteile im Ausland geboren wurden. Alle anderen Schüler/innen gelten als *einheimisch*.

Das für diese Studie gewünschte Messinstrument zur Motivation soll nun zum einen psychometrischen Anforderungen genügen. Zum anderen soll es allen oben angeführten Randbedingungen entsprechen, daher insbesondere über eine schülerfreundliche, verständliche und klare Sprache verfügen und für alle Schüler/innen der Sekundarstufe 1 reliable Messergebnisse liefern.

6.2. Bestehende Instrumente

In Kapitel 5.3.1 wurde der Fragebogen zum Lernen beschrieben (F. H. Müller, et al., 2007). Dieser Fragebogen bezieht sich ebenso auf die theoretische Basis der SDT wie das zu entwickelnde Instrument. Allerdings beschränkt er sich auf die Beschreibung der unterschiedlichen Formen extrinsisch regulierten Verhaltens und lässt aufgrund der Fragestellungen die Beantwortung der Frage nach dem *Warum* einer Handlung offen. Aus diesem Grund wurde dieser Fragebogen im Rahmen der Studie auch dazu eingesetzt, wofür er im obengenannten Artikel beschrieben wurde, nämlich zur Erforschung, wie motiviert Schüler/innen überhaupt sind, sich mit Physik als Fach auseinanderzusetzen.

Ein weiteres Instrument zur Motivation, das in der deutschsprachigen Literatur gefunden werden konnte, war das von Berger und Hänze (2004). Im Rahmen einer Untersuchung über die Wirkung des Gruppenpuzzles im Physikunterricht auf die grundlegenden psychischen Bedürfnisse von Schüler/innen der Sekundarstufe 2 stellen die Autoren ein kurzes Messinstrument vor. Ausgehend von der Tatsache, dass Physikunterricht oft wenig motivierend ist oder sogar abgelehnt wird (Muckenfuß, 1995), versuchen die

Autoren auf der einen Seite durch kontextorientierte Aufgaben in sinnstiftenden Zusammenhängen die Motivation der Schüler/innen zu steigern (Berger, 2000; R. Müller, 2006). Auf der anderen Seite wird versucht, durch eine spezielle Unterrichtsmethode – die des Gruppenpuzzles –, anknüpfend an die SDT, die psychischen Grundbedürfnisse zu erfüllen und so die Motivation zu steigern. Dieses Kreuzdesign (Berger, 2000) erlaubt es einerseits, die zwei verschiedenen, thematisierten Inhalte zu vergleichen. Da jede/r Schüler/in beide Unterrichtsmethoden erlebt, ist es andererseits auch möglich, auf intraindividueller Ebene die Gruppenpuzzle-Methode mit herkömmlichem Frontalunterricht zu vergleichen.

Zur Messung der Motivation wurde ein kurzes Messinstrument entwickelt, das innerhalb weniger Minuten zu bearbeiten ist und verschiedene Unterrichtsmethoden und Unterrichtssituationen (lehrerzentrierte ebenso wie solche, wo sich die Lehrkraft gestaltend im Hintergrund hält) erfassen kann. Die drei Skalen, entsprechend den drei psychischen Grundbedürfnissen der SDT *Soziale Eingebundenheit*, *Kompetenzerleben* und *Autonomieerleben,* wurden mit jeweils zwei Items abgebildet. Diese Items basierten ihrerseits auf einer Adaption der Skalen von Prenzel et al. (1993).

Wie auch eine E-Mail Korrespondenz mit einem Autor der Studie, Prof. Berger (2011), zum Gruppenpuzzle deutlich machte, lag der Fokus bei diesen Skalen darauf, möglichst kurz und universell einsetzbar zu sein. Die Reliabilitäten der Skalen waren mit Werten für Cronbach's α zwischen 0,62 und 0,79 für die Autoren dem Zweck entsprechend ausreichend. Für die Zwecke der hier beschriebenen Studie sollte jedoch ein feineres Instrument mit einer höheren Reliabilität zur Verfügung stehen, damit eine genauere Beschreibung von Motivation und Ergebnissen im Wissenstest möglich ist.

Darüber hinaus ist im Artikel von Berger et al. (ebd.) die Faktorstruktur dieser drei Skalen angegeben. Dabei zeigt sich, dass ein Item der Skala *Autonomieerleben* mit 0,59 eine eher mäßige Ladung auf die Skala besitzt, wenn man die hohe Querladung von ebenfalls 0,59 auf die Skala *Soziale Eingebundenheit* dazu in Relation setzt. Es sei aber darauf hingewiesen, dass die hier beschriebene Studie, aufgrund ihres Designs (Praetest-Posttest – Design), wesentlich genauere Instrumente benötigt, um praktisch bedeutsame Effekte darstellen zu können, als die Studie der Autoren Berger et al. In

ihrer Studie konnten sie aufgrund ihres Testdesigns (Kreuzdesign) durch Wechseln der Gruppen und Treatments die Randbedingungen wesentlich besser kontrollieren.

Aus diesen zwei genannten Gründen wurde darauf verzichtet, dieses Messinstrument für die Zwecke der hier angeführten Studie zu übernehmen. Darüber hinaus geben die Autoren Berger et al. an, dieses Messinstrument für Schüler/innen der Sekundarstufe 2 konstruiert und es an gymnasialen Schüler/innen der Jahrgangsstufe 12 getestet zu haben. Über Erfahrungen mit Jüngeren wird nicht berichtet. Im Gegensatz dazu sollten hier erheblich jüngere Schüler/innen der Sekundarstufe 1 getestet werden, den Jahrgangstufen 5 bis 8 entsprechend, die darüber hinaus Großteils aus einer anderen Schulform, nämlich der Hauptschule stammten. Ein Übernehmen dieses Instruments ohne ein vorheriges Testen auf seine Eignung für die spezielle Stichprobe schien daher nicht angebracht.

Ebenfalls in der deutschsprachigen Literatur zu finden ist die *Kurzskala intrinsischer Motivation (KIM)* von Wilde et al. (2009). Im Kontext von Lernen an außerschulischen Lernorten, wie im berichteten Fall im Berliner Museum für Naturkunde, sollten Aussagen die intrinsische Motivation der Schüler/innen getroffen werden. Außerschulische Lernorte sollen die Neugierde und das Interesse von Schüler/innen wecken und dabei motivationale Lernziele unterstützen. Da die hier angesprochenen Bereiche Interesse, Spaß, die eigene Autonomie im Lernprozess und das Erleben der eigenen Kompetenz sind, führen die Autoren aus, dass die SDT auch für ihre Zwecke den geeigneten theoretischen Rahmen bildet, da Autonomieerleben und Kompetenzerleben zentrale Säulen der SDT sind. Darüber hinaus ist sie auf eine derart breite Menge an Forschungsfeldern anwendbar wie kaum eine andere Theorie zur Motivation. Zur Messung der Motivation konstruierten Deci und Ryan das IMI (2003), das auf der zitierten Webpage in verschiedenen Varianten veröffentlicht ist. Wilde et al. bedienten sich der Standardvariante des IMI, die aus 22 Items besteht, übersetzten diese und bildeten daraus ihre KIM. Die KIM stellte eine verkürzte Version des IMI dar, die aus 12 Items besteht. Ebenso wie die ursprüngliche Version bildet die KIM vier, faktoranalytisch trennbare Skalen mit je drei Items ab: *interest/enjoyment, perceived competence, perceived choice* und *pressure/tension*. Das Antwortformat stellt eine fünfstufige

Likertskala dar. Die Skala *interest/enjoyment* soll nach Angabe der Autoren Deci und Ryan das Konstrukt der intrinsischen Motivation abbilden.

Die Autoren der KIM geben an, dass ihr Messinstrument bei konfirmatorischer Faktorenanalyse zuverlässig die vier Skalen abbildet. Die Reliabilitäten der einzelnen Skalen liegen zwischen 0,53 (*pressure/tension*) und 0,89 (*interest/enjoyment*). Die KIM verfügt über eine ausreichende Retest-Reliabilität.

Dennoch ist bei genauerer Betrachtung einzuwenden, dass die Übersetzungen aus dem Englischen teilweise problematisch sind: So kann z.B. Item 8: „Bei der Tätigkeit in der Ausstellung konnte ich wählen, wie ich es mache."(Wilde, et al., 2009, S 45) zwar der Skale *perceived competence* zugeordnet werden, aber innerhalb dieser Skala ist es bei wörtlicher Übersetzung schwer, es einem der Originalitems zuzuordnen. Item 12 aus der KIM lautet „Ich hatte Bedenken, ob ich die Tätigkeit in der Ausstellung gut hinbekomme" (ebd.). Dieses Item wird als Teil der *pressure/tension* Skala angegeben. Abgesehen davon, dass das keine quellennahe Übersetzung eines der originalen Items aus dieser Skala ist, stellt es sich auch bei der Faktorenanalyse als problematisch heraus und würde aufgrund der Übersetzung inhaltlich eher zu Skala *perceived competence* passen.

Von der Altersstufe her, an der dieses Instrument getestet wurde (Jahrgangsstufe 5), scheint die KIM hingegen genau auf die Zwecke der hier berichteten Studie zu passen.

Bezüglich der Faktorladungen der KIM werden von den Autoren nur Faktoren angegeben, die größer als 0,50 sind oder kleiner als -0,50 sind. Während die meisten Items Ladungen zwischen 0,64 und 0,90 haben, ergeben sich für Item 1 und Item 3 eher als mäßig zu bewertende Ladungen von 0,51, bzw. 0,56 (Nachtest I), überhaupt wenn man bedenkt, dass es Querladungen bis 0,5 geben könnte, die hier nicht angegeben sind. Item 12 zeigt eine bestmögliche Faktorladung (Nachtest II) von 0,61, im Nachtest I sogar eine hohe negative Querladung (-0,68) auf eine andere Skala (*interest/enjoyment*). Das ist, je nachdem, wie groß die restlichen Ladungen sind, über die hier keine Angaben gemacht wurden, als nicht übermäßig gut zu bewerten.

Der Grund, warum die KIM nicht übernommen wurde, liegt aber hauptsächlich darin, dass sie eine, im Kontext des CAPT-Prozesses ganz wesentliche Skala nicht berücksichtigt: die *effort/importance*-Skala, also im weitesten Sinne eine Skala, die

Anstrengungsbereitschaft für etwas, was einem selbst wichtig erscheint, testet. *Effort/importance* scheint ein wesentliches Konstrukt zu sein, wenn man, wie in unserem Fall, evaluieren will, wie erfolgreich CAPT als Intervention ist. Denn es ist davon auszugehen, dass engagiertere Tutoren, die die Wichtigkeit es eigenen Tuns erkennen, bessere Ergebnisse für sich und ihre Tutees erzielen werden. Daher kann im Kontext von CAPT auf diese Skala nicht verzichtet werden. Einem Messinstrument, das diese Skala nicht berücksichtigt, fehlt für die intendierte Beschreibung daher ein wesentlicher Faktor.

Hingegen berücksichtigt die KIM mit der Skala *pressure/tension* etwas, was im Kontext des CAPT – Prozesses nicht sehr aussagekräftig erscheint, da die Teilnahme beim Tutoring weder auf Druck von außen erfolgte (die Teilnahme konnte auch abgelehnt werden) noch zur Notengebung herangezogen wurde. Darüber hinaus ergeben Studien (Kim & Gill, 1997; Whitehead & Corbin, 1991), dass die *pressure/tension*-Skala hinsichtlich ihrer Reliabilität inakzeptabel schlechte Werte liefert. Die Autoren dieser beiden Studien verzichten daher gänzlich darauf, diese Skala zu verwenden.

Da keines der verfügbaren Instrumente allen Anforderungen entsprach, wurde beschlossen, das IMI in die deutsche Sprache zu übersetzen und an die zu untersuchende Stichprobe zu adaptieren.

6.3. Konstruktion eines Messinstrumentes aus dem IMI

Als Basis des hier vorgestellten Messinstrumentes zur Motivation wurde das Konstrukt der Selbstbestimmungstheorie, der SDT (Deci & Ryan, 1985, 1993, 2002) gewählt. Dazu passend stellen die Erfinder der SDT ein Messinstrument, das Intrinsic Motivation Inventory (IMI) zur Verfügung (Deci & Ryan, 2003). Dieser Fragebogen wurde für die Zwecke dieser Studie aus dem Amerikanischen übersetzt und bearbeitet.

6.3.1. Die Skalen des IMI

Das IMI versteht sich als multidimensionales Messinstrument, das bereits in vielen Situationen im Zusammenhang mit intrinsischer Motivation eingesetzt wurde. Es besteht aus sieben Subskalen und beinhaltet in seiner Vollversion 45 Items, das sind fünf bis acht

Items pro Skala, die teilweise redundant sind. Den Ergebnisraum bildet eine siebenteilige Likert-Skala.

Die sieben Subskalen sind:

- perceived competence – (selbst) wahrgenommene Kompetenz
- perceived choice – wahrgenommen Wahlfreiheit
- interest and enjoyment – Interesse und Vergnügen
- effort and importance – Einsatz und Wichtigkeit
- value and usefulness – Nützlichkeit und Wert
- (social) relatedness - Eingebundenheit
- felt pressure and tension – wahrgenommener Druck und Anspannung

Die *interest/enjoyment*-Skala ist jene, die nach Abgaben der Autoren die intrinsische Motivation selbst abbildet. Die Skala liefert eine Selbsteinschätzung in wie weit Handeln intrinsisch motiviert ist. Es ist die einzige Skala, die sich direkt auf intrinsische Motivation bezieht und besteht daher aus mehr Items als die restlichen Skalen.

Ergänzend dazu sind die beiden Konstrukte *perceived competence* und *perceived choice* positive Prädiktoren für das Selbstkonzept und die intrinsische Motivation einer Person. Hingegen bildet die Skala *pressure/tension* einen negativen Prädiktor für intrinsisch motiviertes Handeln.

Die *effort/importance*-Skala bildet ein eigenes Konstrukt ab, das sich nicht direkt auf *intrinsische* Motivation bezieht. Sie beschreibt vielmehr die Bereitschaft, sich für etwas anzustrengen, das als wichtig erkannt wurde, also man einen bestimmten Zweck verfolgen will. Im Sinne vieler moderner Motivationstheorien spiegelt eine derartige Bereitschaft motiviertes Verhalten wider (Deci & Ryan, 1993). Diese Skala quantifiziert daher die Ausprägung an Motivation, sie beschreibt das *Wieviel* an Motivation einer Person, während die restlichen Skalen die Ursachen und den Grad der Selbstbestimmtheit angeben.

Die Skala *value/usefulness* kann immer dann verwendet werden, wo es um Internalisierung geht. Denn wenn Personen erkennen, dass ihre Handlungen brauchbar und nützlich sind, werden sie die zugrunde liegenden Haltungen eher verinnerlichen und selbstgesteuertes Verhalten an den Tag legen.

Die *relatedness*-Skala kann in Situationen Anwendung finden, wo Interaktionen von Personen beschrieben werden sollen. Sie versucht zu erfassen, inwieweit es zu einer Verbundenheit der Agierenden kommt, inwieweit es zu einem Gefühl der sozialen Zugehörigkeit kommt. Deci und Ryan (1993) gehen davon aus, dass Personen, die sich zu einer Gruppe zugehörig fühlen, daran interessiert sind, innerhalb dieser Gruppe etwas zu bewirken.

Die drei Skalen *interest/enjoyment, pressure/tension und effort/importance* sind somit Indikatoren für intrinsisch motiviertes Verhalten (Markland & Hardy, 1997).

Deci und Ryan geben auf ihrer Webpage an (2003), dass die Items, die diese Skalen bilden, kohärent und faktoranalytisch trennbar sind. Die Skalen sind für eine große Anzahl von Anwendungsgebieten einsetzbar und stabil. Damit ein Item in eine Skala inkludiert wurde, musste es eine Faktorenladung von mindestens 0,6 auf diese Skala aufweisen und durfte keine Querladung über 0,4 (bzw. unter –0,4) haben, was für eine Vielzahl von Items bei weitem zutrifft. Weiters empfehlen die Autoren für den Einsatz des IMI in eigenen Forschungsprojekten lediglich jene Skalen zu verwenden, die für den persönlichen Zweck relevant sind. Sie begründen dies damit, dass bislang keine Effekte der Itemordnung nachweisbar sind. Auch ermutigen sie explizit dazu, die Items den spezifischen Aktivitäten anzupassen und haben dafür Vorkehrungen getroffen: In der *value/usefulness*-Skala lautet Item 2 im Originaltext: „*I think that doing this activity ist useful for* _____“. Hier ist bereits die Ergänzung des eigenen Kontextes vorgesehen, woraus sich keine Effekte auf die Reliabilität oder die Validität des Fragebogens ergeben sollten.

Darüber hinaus ist vorgesehen, dass Items, die redundant erscheinen, weggelassen werden können. Die Empfehlung lautet hier, dass etwa vier Items eine Skala ergeben. Jedes weitere Item würde lediglich zu einer kleinen Verbesserung der Varianzaufklärung beitragen.

6.3.2. Psychometrische Eigenschaften des IMI

Bezüglich der Validität des Fragebogens geben die Autoren (Deci & Ryan, 2003) an, dass die Items „*...quite face-valid*“ sind.

Hinsichtlich der Reliabilität ihres Fragebogens machen die Autoren zunächst keine Angaben. In persönlicher Korrespondenz schrieb Prof. Deci: „*I don't know about its reliability, but I feel sure it would be high*" (2014).

Vielleicht mahnen die Autoren aufgrund dieser relativ schwachen Aussagen zu Validität und Reliabilität zu Vorsicht bei der Interpretation der mit dem IMI erhobenen Daten. So zeigt sich, dass der Fragebogen nur dann ein zuverlässiges Maß für intrinsische Motivation darstellt, wenn die Items der Skala *perceived competence* und *interest/enjoyment* signifikant korreliert sind (Ryan, Koestner, & Deci, 1991).

Leider gibt es hinsichtlich dieser psychometrischen Eigenschaften des IMI wenige Publikationen der Entwickler selbst. Das ist vielleicht auch ein Grund dafür, dass die Autoren Deci und Ryan empfehlen, im eigenen Datensatz konfirmatorische Faktorenanalysen durchzuführen.

Recherchen über das IMI, über die Konstruktion der Skalen und die psychometrischen Eigenschaften in der Literatur ergeben ein etwas differenzierteres Bild, wenngleich die Anwendungsgebiete fast ausschließlich im sportlichen Kontext zu finden sind. Als Beispiel sei die Motivation von Sportler/innen zu teilweise hartem Training genannt.

Ryan (1982) erklärt für die *effort/importance* Skala, dass die Items ad-hoc Konstrukte sind. Auch Markland et al. (1997) geben betreffend der Herkunft des IMI[12] zunächst an, dass Ryan (1982) üblicher Weise zitiert wird. Einschränkend muss aber erwähnt werden, dass in diesem Artikel lediglich über die drei Skalen *interest/enjoyment*-, *pressure/tension*- und *effort/importance*, jeweils als 7-teilige Likertskalen mit insgesamt 26 Items, berichtet wird. Diese werden im Kontext von Puzzle-Lösen angewandt, um den Einfluss unterschiedlicher Feedbackformen auf die intrinsische Motivation zu beschreiben. Von Ryan et al. (1983) sind in diesem Zusammenhang auch faktoranalytische Berechnungen zu diesen drei Skalen publiziert: *interest/enjoyment* mit 11 Items, *pressure/tension* mit 3 Items und *effort/importance* mit lediglich 2 Items. In diesem Artikel wird der Fragebogen aber nicht als IMI bezeichnet, obwohl es sich offensichtlich um dasselbe Instrument handelt.

[12] Markland (1997, p.21) schreibt hier wörtlich: „The origins of the IMI, then, are somewhat shrouded in mystery."

Es gibt zwei weitere Publikationen zu den psychometrischen Eigenschaften des IMI: (McAuley, Duncan, & Tammen, 1989; McAuley, Wraith, & Duncan, 1991). Beide Untersuchungen finden in einem sportbezogenen Kontext statt, betreffend Wettkampfsportler und ihre Motivation, sich dem teilweise harten Training zu unterziehen. Hier wird bereits von einem 5-dimensionalen Instrument gesprochen, das zusätzlich zu den drei oben erwähnten Skalen noch *perceived competence* und *perceived choice* als Skalen erwähnt. Über die letzte Skala wird gesagt, dass sie kurz zuvor hinzugefügt wurde, aber noch nicht validiert ist. McAuleys Bestreben die Faktorstruktur des IMI in dieser Form zu bestätigen, liefert Ergebnisse, die nicht gerade zur Euphorie beitragen: Zum Beispiel lädt das Item, das als Item 14 bezeichnet wird, mit 0,46 bis 0,47, je nach berechnetem Modell, nur sehr mäßig auf die erwünschte Skala (*perceived competence*). Querladungen werden keine angegeben. McAuley schließt daraus, dass die Messung des Konstrukts *Intrinsische Motivation* weniger weit entwickelt ist als das Konstrukt selbst.

Markland und Hardy (1997) untersuchen anhand von 169 Proband/innen die Faktorstruktur und die Validität des IMI. Dabei beziehen sie sich ebenfalls auf einen sportbezogenen Kontext, indem sie Sportstudierende befragen. Im Gegensatz zu den oben zitierten Studien liegt bei ihnen der Forschungsfokus auf dem wahrgenommenen Sitz der Entscheidungen (*perceived locus of causality*), also ob Entscheidungen von außen oder aus dem Individuum selbst kommen, und weniger auf der *perceived competence*. Die Autoren erachten das Konzept des *perceived locus of causality* als zentral in der SDT und entwickelten dazu fürs erste ein Instrument, das den Sitz der Entscheidungen erfassen sollte. Faktorenanalysen ergeben Skalen mit konsistenten Items, denen offensichtliche eine hohe Validität *(high face validity*) bescheinigt wird. Die in dieser Studie untersuchten Skalen des IMI waren *interest/enjoyment, pressure/tension, effort/importance* und *perceived competence*. Während die ersten drei Konstrukte Indikatoren für intrinsische Motivation sind, ist die *perceived competence* gemeinsam mit der hier neu entwickelten Skala für den *locus of causality* auf einer anderen konzeptuellen Ebene angesiedelt. Zusammenfassend geben die Autoren als Ergebnis der unterschiedlichen Strukturgleichungsmodelle, die hier gerechnet wurden, an, es wäre argumentierbar ist, dass sich Personen zwar kompetent für eine Aktivität

fühlen, aber daraus nicht notwendiger Weise folgt, dass sie dafür auch intrinsisch motiviert sind diese durchzuführen. Kritisiert wird insbesondere die *effort/importace*-Skala, weil auch für Belohnungen von außen (z.B. Bezahlung) große Anstrengungen unternommen würden und diese Skala daher nicht unbedingt die Anstrengung für intrinsisch motivierte Tätigkeiten widerspiegle.

Dem IMI als solchem werden mögliche konzeptuelle und operationale Schwächen bescheinigt. Die Autoren empfehlen daher, die Konstruktvalidität des IMI zu prüfen, bevor es eingesetzt wird.

Eine neuere Publikation, die sich auch mit der Faktorstruktur des IMI auseinandersetzt, stammt von Tsigilis und Theodosiou (2003). Zwar liegt der Hauptzweck der Studie darin, die Test-Retest-Reliabilität des IMI zu überprüfen. Als Basis dafür wurden jedoch auch die Konstruktvalidität der griechischen Version des IMI und eben die Faktorstruktur des IMI geprüft. Auch hier war der Kontext der Untersuchung ein sportlicher: 144 Studierende unterzogen sich freiwillig einer Austestung ihrer aeroben Schwelle beim Ausdauertraining. Die Autoren beziehen sich für ihre Untersuchungen auf das IMI in der Version, die auch schon McAuley et al. (1989) verwendeten, benutzten aber nur die vier Skalen *interst/enjoyment, effort/importance, perceived competence* und *pressure/tension* mit jeweils vier bzw. fünf Items. Während die internen Korrelationen (Cronbach's α) der ersten drei Skalen gut sind (0,78, 0,84 und 0,80), ist jene der *pressure/tension* Skala mit 0,66 schlechter, weshalb diese Skala auch in den weiteren Analysen nicht berücksichtigt wurde. Die drei verbleibenden Skalen erklären 61,0 % der Varianz. Was die Test-Retest-Reliabilitäten der einzelnen Skalen anbelangt, so sind jene der Subskalen bei weitem nicht akzeptabel, während der Fragebogen als Ganzes einen akzeptablen ICC (Intraclass Correlation Coefficient) von 0,70 aufweist. Die Autoren können einen Mediationseffekt der *perceived competenc*-Skala zur intrinsischen Motivation feststellen, was im Einklang mit früheren Studien steht (Vallerand, 2000; Whitehead & Corbin, 1991).

Probleme mit der Subskala *pressure/tension* werden auch von den Autoren Kim und Gill (1997) berichtet. Ihre Untersuchungen an 344 Mittelschüler/innen (etwa 14 Jahre alt) verwendeten ebenfalls die Items entsprechend der vier Subskalen von McAuley: *interst/enjoyment, effort/importance, perceived competence* und *pressure/tension*.

Explorative wie auch konfirmatorische Faktoranalysen, die zum Ausschluss einzelner Items führten, zeigten Reliabilitäten (Cronbach's α) der *pressure/tension*-Skala von 0,57, bzw. 0,62, je nach inkludierten Items. Das sind für die Autoren inakzeptable Werte. Auch die Korrelation dieser Skala zum IMI als Ganzes war gering. Daher wurde diese Skala auch in dieser Studie von weiteren Analysen ausgeschlossen. Hingegen ergaben die Cronbach's α für die restlichen drei Skalen zufrieden stellende Werte (von 0,72 bis 0,80).

Zusammenfassend ist festzustellen, dass das IMI vor allem im Kontext sportlicher Aktivitäten, Mannschaftssport, wie auch Einzelsport, getestet und verwendet wurde. Für die Eignung des IMI in anderen Kontexten gibt es wenige Evidenzen, abgesehen von zwei Publikationen (Ryan, 1982; Ryan, Connell, & Robert, 1990). Bedauerlich ist, dass über die Kontexte, in denen die einzelnen IMI-Items konstruiert wurden, nicht berichtet wird. Deci und Ryan (2003) geben zwar an, dass ihr Instrument in den unterschiedlichsten Situationen angewandt werden kann, jedoch fehlen Dokumentationen der nötigen Studien dazu.

6.3.3. Übersetzung und Re-Übersetzung des IMI

Ausgehend davon, was in der Literatur über das IMI zu finden war, wurde für die hier berichtete Studie dieses Instrument in seiner vollen Version, alle Subskalen und Items einschließend, zur Hand genommen. Die Ausnahme bildete die Subskala *pressure/tension*, sie wurde ausgeschieden, da sie schon aus theoretischen Überlegungen nicht in den zu untersuchenden Kontext passte (vgl. 6.2). Darüber hinaus zeigten sich in der Vergangenheit bei der Verwendung dieser Skala die oben berichteten psychometrischen Schwierigkeiten.

Die verbleibenden 40 Items des IMI wurden von der Autorin dieses Textes aus dem amerikanischen Original übersetzt. Danach wurden sie einer, an dieser Forschung sonst unbeteiligten Anglistin zur unabhängigen Re-Übersetzung vorgelegt. Im Zuge des Vergleichs der originalen Items mit den re-übersetzten wurden einige Formulierungen in der deutschen Version adaptiert. Es fand ein Prozess des Aushandelns einer adäquaten Übersetzung statt, da diese unterschiedlichen Ansprüchen genügen sollte. Zum einen beinhaltete dieser Prozess die Rücksichtnahme darauf, dass die Zielgruppe, die diese Items beantworten sollte, 10 bis 14-Jährige waren. Daher wurde versucht, eine schülernahe Sprache zu finden, die aus der Perspektive von Schüler/innen ansprechend

erschien und verständlich, aber trotzdem nicht zu umgangssprachlich war. Es wurde auf Modeworte der Jugendsprache (wie z.B. „cool“) verzichtet. Zum anderen war vom Design der Studie her von vornherein klar, dass nicht nur sprachlich eher versierte Gymnasialschüler/innen diesen Fragebogen zu beantworten hätten, sondern durchschnittliche bis kognitiv eher schlechte Schüler/innen aus Hauptschulen sowie Schüler/innen mit Migrationshintergrund im Zentrum der Untersuchungen stehen würden. An eben diese Schüler/innen sollten die Formulierungen der Items von der Verständlichkeit, der Satzlänge und dem verwendeten Vokabular her angepasst sein.

Das Ergebnis dieses Aushandlungsprozesses ist Teil der Validierung des Fragebogens und ist im Anhang 13.1 zu finden. Mit diesen Items wurde eine erste Pilotierung gestartet, um redundante Fragestellungen entfernen und konfirmatorisch die Faktorstruktur des IMI reproduzieren zu können.

6.4. Erste Pilotierungen, auftretende Schwierigkeiten und zweite Pilotierung

Der im Rahmen dieser Studie entwickelte Fragebogen zur Motivation, der aus dem IMI entstand, bestand nun aus sechs Skalen (*perceived competence, perceived choice, interest and enjoyment, effort and importance, value and usefulness, (social) relatedness*) und insgesamt 40 Items (siehe Anhang 13.2).

Als Ergebnisraum wurde eine fünfteilige Likert-Skala gewählt mit 1 = „stimmt gar nicht“ bis 5 = „stimmt völlig“. Im Unterschied zum ursprünglichen IMI, das mit einer siebenteiligen Likert-Skala arbeitet, wurde im Kontext mit Schüler/innen der Sekundarstufe 1 eine fünfteilige Skala als ausreichend empfunden. Da Validität und Reliabilität mit der Anzahl der Antwortkategorien zunächst steigen, wird zwar eine siebenstufige Skala oft als Optimum empfohlen (Preston & Colman, 2000). Allerdings können die psychometrischen Eigenschaften eines Tests darunter leiden, wenn zu viele Antwortkategorien vorhanden sind. Das kann dann der Fall sein, wenn die Proband/innen mit dem Differenzierungsgrad überfordert sind (Bühner, 2011). Da die Items des IMI von der Formulierung her das Konstrukt sehr differenziert abfragen, die Schüler/innen mit 10 bis 14 Jahren jung sind und oft eine andere Erstsprache als Deutsch

aufweisen, wurde aus Gründen der verbalen Verständlichkeit die fünfstufige Variante gewählt.

Der Fragebogen zur Motivation wurde sowohl in einer Online-Version, als auch einer Paper-Pencil-Version, an N = 238 Schüler/innen aus 10 verschiedenen Klassen der Sekundarstufe 1 pilotiert. Die Schüler/innen nahmen freiwillig und anonym an dieser Pilotierung teil und waren nicht ident mit jenen, die später im CAPT-Projekt mitmachten. Es wurde auf eine möglichst breite Streuung hinsichtlich des Migrationshintergrundes und der Art der Schule (Allgemein bildende höhere Schule oder Hauptschule) geachtet. Dennoch stammten die Schüler/innen teilweise aus denselben Schulen wie die CAPT-Projektklassen, was aber aufgrund der breiten Streuung zu keiner Einschränkung der Aussagekraft führt.

Die Instruktion an die durchführenden Lehrer/innen war, dass sie den Fragebogen nach einer Gruppen- oder Partnerarbeit einsetzen sollten, unabhängig vom Fach oder von einem spezifischen Kontext. Der Grund für die empfohlene Sozialform war der, dass die Fragen nach der sozialen Eingebundenheit (*relatedness*-Skala) für die Schüler/innen sonst keinen Sinn ergeben hätten.

In einem einleitenden Text wurden die Schüler/innen darauf hingewiesen, dass sie durch das Ausfüllen des Fragebogens einer Forscherin an der Universität helfen sollen, den Fragebogen zu testen. Die Wörter „Universität“ und „forschen“ sollten vermitteln, dass die Schüler/innen in einem wichtigen Bereich um ihre Expertise befragt werden und daher den motivationalen Bereich *perceived competence* ansprechen, somit zu einer höheren Motivation beim Ausfüllen beitragen, was wiederum exaktere Ergebnisse liefern sollte. Auch die Problematik der redundanten Items wurde angesprochen und den Schüler/innen erklärt, dass mit Hilfe ihrer Antworten die geeignetsten Formulierungen gefunden werden sollten.

Da diese erste Pilotversion des Motivationsfragebogens in unterschiedlichsten Kontexten eingesetzt wurde, war es in der Ausformulierung der Items nicht möglich, auf die spezifische Situation einzugehen. Es wurde daher innerhalb der Items eher allgemein von „dieser Tätigkeit“ gesprochen. Wenn man von der Empfehlung von Busker (Busker, 2014, p.273) ausgeht, Items möglichst konkret und direkt an Sachverhalte anzupassen, stellt das eine große Abstraktion dar, die von den Schüler/innen gefordert wurde.

Die Daten aus dieser ersten Pilotierung wurden mittels konfirmatorischer Faktorenanalyse untersucht, um die Skalenstruktur des IMI reproduzieren zu können. Dies war leider nicht möglich. Im Folgenden wird versucht, einzelne Bereiche zu identifizieren, die als mögliche Ursachen für diese Schwierigkeiten infrage kamen.

6.4.1. Verständnisprobleme

Trotz sorgfältiger Übersetzungen und trotz Expertenvalidierung der übersetzten Items durch zwei weitere Lehrkräfte ergaben sich Probleme mit dem Verständnis von einzelnen Wörtern, die in einzelnen Items verwendet wurden.

Dazu ausgewählte Beispiele:

Das erste Originalitem aus der *relatedness*-Skala des IMI lautet:

I felt really distant to this person.

Es wird in weiterer Folge als rel 1 bezeichnet. Damit sind die Subskala und Itemnummer klar bezeichnet.

Es wurde übersetzt mit:

Ich fühlte mich der Person richtig fern.

In Übereinstimmung berichteten acht befragte Schüler/innen aus der 2. Klasse der Hauptschule, dass sie das Wort „fern“ in diesem Zusammenhang nicht verstanden. Weil dieses Item eine andere, negative Polung hatte, darüber hinaus unverständlich war und weil es keine guten Korrelationen (zwischen 0,176 und 0,358 in zwei Fällen 0,525 und 0,536, jeweils mit $p < 0,01$) zu den restlichen Items dieser Skala hatte, wurde es von weiteren Analysen ausgeschlossen.

Am Beispiel des achten Items (rel 8) aus derselben Skala wird eine andere Vorgangsweise demonstriert. Das Originalitem lautet:

I feel close to this person.

Der erste Versuch einer expertenvalidierten Übersetzung lautete:

Ich fühle mich dieser Person nahe.

Auch hier ergab ein Nachfragen bei denselben Schüler/innen, dass das Wort „nahe“ nicht verstanden wurde. Bei diesem Item wurde der Versuch gemacht es durch eine dem Original zwar textfernere, aber angemessenere Formulierung zu reparieren:

Ich konnte spüren, wie es dem anderen Schüler / der anderen Schülerin ging.

Aufgrund der Antworten und Einwände dieser acht Schüler/innen wurden einige Items umformuliert, andere wiederum gestrichen. Weil sich aus Analysen zu negativ gepolten Items ergab, dass diese große Verständnisschwierigkeiten auslösten, wurden Items dieser Art gänzlich aus dem Fragebogen entfernt.

6.4.2. Probleme mit der Formulierung

Unter *Problemen mit der Formulierung* werden im Wesentlichen zwei Aspekte subsummiert. Zum einen unterscheiden die Items des IMI eine Feinheit an Empfindungen, die von den Schüler/innen so nicht wahrgenommen wurde. Entweder differenzierten die Schüler/innen nicht so detailliert in ihren Empfindungen, oder die sprachliche Formulierung der Unterschiede wurde nicht verstanden.

Dazu ein Beispiel: Die zwei Originalitems aus der *relatedness*-Skala rel 4 und rel 7 lauten:

I felt I had to do this – Mir kam vor, dass ich das tun musste.

und

I did this activity because I had to – Ich machte diese Aktivität, weil ich musste.

Hier wurde von Schüler/innen berichtet, dass sie keinen Unterschied zwischen den Items erkennen können. Dies ist entgegen der offensichtlichen Intention des IMI, da der Unterschied zwischen beiden Formulierungen im Ort der Entscheidung liegt (*locus of causality*). Im ersteren Item (rel 4) ist dieser in der Person selbst zu finden, im zweiteren (rel 7) irgendwo außerhalb, bei einer Person, die Zwang ausübt.

Aus derselben Skala stammen zwei Items (rel 5 und rel 6), die in der ursprünglichen Version des IMI offensichtlich widersprüchliche Aussagen haben sollten:

I did this activity because I had no choice – Ich machte hier mit, weil ich keine andere Wahl hatte.

und

I did this activity because I wanted to – Ich machte hier mit, weil ich es wollte.

Schüler/innen empfanden hier keinen Widerspruch. Wenn man die intendierte negative Polung des ersteren Items berücksichtigt, sollten beide Items hochkorreliert sein und auf dieselbe Skala laden. Dem ist aber nicht so. Im Schulalltag ist vielleicht zu argumentieren, dass Schüler/innen dieser Widerspruch nicht bewusst ist oder er ihnen ausgeredet wird: Als schulpflichtige Schüler/innen haben sie keine andere Wahl als in die

Schule zu gehen – also *no choice*. Dem gegenüber wird oft genug von Seiten der Erziehenden betont, dass sie beim Lernen ein gewisses Maß an *Wollen* zeigen *müssen* (wie widersprüchlich!), damit es zu fruchtbaren Ergebnissen führt. Darüber hinaus ist zu hinterfragen, ob es bei Schüler/innen dieses Alters überhaupt einen Widerspruch zwischen „etwas tun müssen“ und „etwas gerne tun“ gibt. Das würde bereits die Existenz intrinsischer Motivation voraussetzen, die aber in diesem Alter unter Umständen noch nicht voll entwickelt ist, einfach weil die autonomen Persönlichkeitsstrukturen noch fehlen (Vallerand, 2000).

6.4.3. Probleme der Itempolarität

Probleme mit der Polarität der Items ergaben sich durch alle Skalen hindurch. Anhand der *effort/importance*-Skala werden sie im Folgenden genauer beschrieben.

Für positiv gepolte Items bedeutet eine hohe Zustimmung eine hohe Ausprägung des Merkmals. Negativ gepolte Items werden vor allem dazu verwendet, um einer Zustimmungstendenz (Akquieszenz) der Befragten entgegenzuwirken. Denn oft ist es so, dass man leichter einer Formulierung zustimmt (z.B. „Ich bin glücklich“) als deren Gegenteil („Ich bin unglücklich“) ablehnt (Bühner, 2011). Um eben dies zu vermeiden, werden Items mit entgegengesetzter Polung verwendet.

Die *effort/importance*- Skala beinhaltet in der ursprünglichen Version des IMI fünf Items. Diese lauten:

eff 1: *I put a lot of effort into this.*
Ich habe mich sehr eingesetzt bei dieser Tätigkeit.

eff 2: *I didn't try very hard to do well at this activity.* (R)
Ich habe mich nicht sehr angestrengt diese Tätigkeit gut zu machen.

eff 3: *I tried very hard on this activity.*
Ich habe mich sehr angestrengt bei dieser Tätigkeit.

eff 4: *It was important to me to do well at this task.*
Es war wichtig für mich, diese Aufgabe gut zu bewältigen.

eff 5: *I didn't put much energy into this*. (R)
Ich habe nicht viel Energie hineingesteckt.

Die Items eff 2 und eff 5 sind mit einem (R) markiert und daher inhaltlich invers zu den restlichen Items. Kreuzt man hier „stimmt völlig“ an, arbeitet man nicht mit vollem Einsatz. Im Gegensatz dazu bedeutet ein „stimmt völlig“ bei den Items eff 1, eff 3 und eff 4 volle Anstrengungsbereitschaft. Man spricht daher hier von einer negativen Polung der Items eff 2 und eff 5.

Für die bereits umkodierten[13] Items wurden zunächst die Histogramme der einzelnen Items begutachtet:

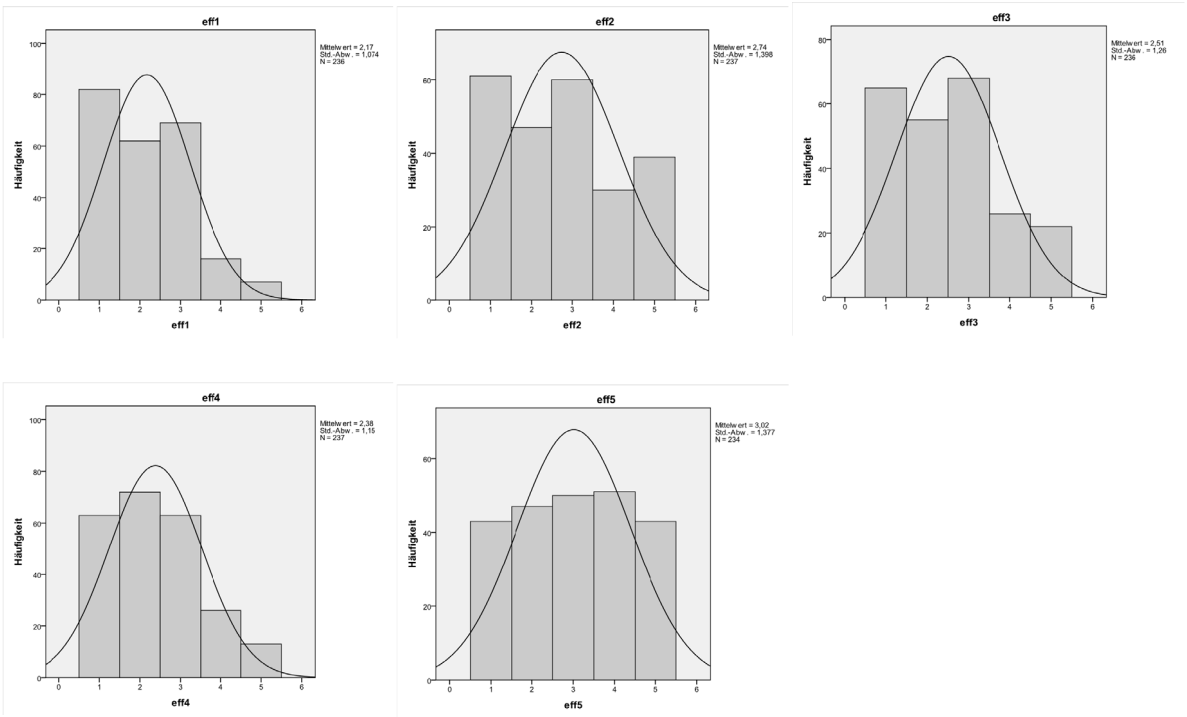


Abbildung 6.1: Histogramme der fünf Items aus der *effort/importance*-Skala

Es ergeben sich optisch wenig Auffälligkeiten, außer, dass links- bzw. rechtssteile Verteilungen zu beobachten sind. Eff 2 ist im Gegensatz zur Erwartung nicht rechtssteil, obwohl es eine negative Polung gegenüber der Items eff 1, 3 und 4 aufweist. Item 5 hingegen ist leicht rechtssteil.

Tabelle 6.1 gibt die bivariaten Pearson-Korrelationen zwischen den einzelnen Items der *effort/importance*-Skala an. Man kann erkennen, dass die beiden ursprünglich negativ gepolten Items (eff 2 und eff 5) miteinander hoch korrelieren, während die restlichen

[13] Der umkodierte, neue Wert ergibt sich mit 6 – (alter Wert).

drei Items miteinander korrelieren. Korrelationen zwischen positiv und negativ gepolten Items fallen niedrig aus. Nach dem Umkodieren sollten jedoch alle Items untereinander hoch korreliert sein.

Inter-Item-Korrelationsmatrix					
	eff1	eff2	eff3	eff4	eff5
eff1	1,000	0,070	0,446	0,303	0,179
eff2	0,070	1,000	0,019	0,076	0,517
eff3	0,446	0,019	1,000	0,401	0,119
eff4	0,303	0,076	0,401	1,000	0,057
eff5	0,179	0,517	0,119	0,057	1,000

Tabelle 6.1: Korrelationen zwischen den einzelnen Items der Skala *effort/importance*

Berechnet man auf Basis von Cronbach's α die Reliabilität dieser Skala, fällt sie mit 0,576 nicht zufriedenstellend aus. Das ist aber aufgrund der teilweise geringen Korrelationen nicht anders zu erwarten gewesen.

Für die *effort/importance*-Skala wurde eine explorative Faktorenanalyse durchgeführt. Bei einer solchen wird versucht, für eine Anzahl von n Variablen (Items) möglichst wenige Faktoren (latente Variable bzw. Subskalen) zu finden, die gleichzeitig möglichst viel der Varianz des gesamten Variablensets erklären. Grafisch gesehen werden orthogonale Achsen gesucht, die den n-dimensionalen Ergebnisraum in seiner jeweils maximalen Ausdehnung möglichst gut beschreiben. Der zugehörige Eigenwert eines Faktors gibt den Anteil der aufgeklärten Varianz an allen Variablen an. Ein Eigenwert, der größer als 1 ist, bedeutet, dass dieser Faktor so viel Varianz aufklärt, wie jede der (standardisierten) Variablen aufweist. Das kann als Selektionskriterium für einen Faktor genutzt werden (Kaiser-Gutman-Kriterium). Bei der anschließenden Hauptkomponentenanalyse werden die Faktoren so optimiert, dass die aufgeklärte Varianz maximiert wird (Bühner, 2011).

Für die Skala *effort/importance* ergibt eine explorative Faktorenanalyse mit anschließender Hauptkomponentenanalyse auf Basis des Kaiser-Kriteriums (Eigenwerte >1), dass sie in zwei Subskalen zerfällt. Tabelle 6.2).

Komponente	anfängliche Eigenwerte		
	gesamt	% der Varianz	kumulierte %
1	**1,895**	**37,893**	**37,893**
2	**1,403**	**28,057**	**65,951**
3	0,716	14,324	80,275
4	0,531	10,617	90,892
5	0,455	9,108	100,000

Tabelle 6.2: Faktorenanalyse der fünf Items der Skala *effort/importance*

Die beiden Subskalen (Tabelle 6.3), die bei der Rotation erhalten wurden, entsprechen genau der Polung der Items. Die Skala zerfällt somit in eine Subskala mit Items positiver Polarität (das sind die Items eff1, eff3 und eff4) und einer mit negativer Polarität (Items eff2 und eff5).

	Komponente	
	1	2
eff1	0,749	
eff2		0,873
eff3	0,821	
eff4	0,724	
eff5		0,863

Tabelle 6.3: Rotierte Komponentenmatrix. Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Varimax mit Kaiser-Normalisierung. Werte < 0,2 wurden unterdrückt.

6.4.4. Doppelte Verneinungen

Ein weiteres Problem ergab sich bei der Verwendung negativ gepolter Items durch doppelte Verneinungen, nicht im streng grammatikalischen, aber im inhaltlichen Sinn. Dies sei anhand der *relatedness*-Skala erklärt: Die beiden Items

rel 5 und rel 6 aus dieser Skala lauten im Original und in der Übersetzung:

rel 5: I'd really prefer not to interact with this person.

Ich würde mit dieser Person lieber nicht mehr zusammenarbeiten.

rel 6: I don't feel like I could really trust this person.

Ich habe nicht das Gefühl, dass ich dieser Person wirklich vertrauen kann.

Eine „Expertenvalidierung" mit 8 SchülerInnen der 2. Klasse Hauptschule ergab, dass

6 von 8 Schüler/innen die doppelte Verneinung nicht anwenden konnten. Das bedeutet, sie konnten nicht sagen, ob sie „stimmt völlig" oder „stimmt gar nicht" ankreuzen sollten im Falle, dass sie doch gerne wieder mit einer bestimmten Person zusammenarbeiten wollten oder ihr doch vertrauen konnten.

Wie man der Literatur zur Testkonstruktion nachlesen kann (z.B. Bühner, 2011), ist die Verwendung von Items mit unterschiedlicher Polarität durchaus umstritten. Der wichtigste Grund dafür ist, dass die unterschiedliche Polarität die Faktorstruktur eines Messinstrumentes beeinflussen kann, wie das auch hier der Fall ist. Die Gründe dafür liegen einerseits darin, dass unterschiedlich gepolte Items oft unterschiedliche Mittelwerte aufweisen und man daher nicht davon ausgehen kann, dass beide Itemarten ein Konstrukt auf dieselbe Weise abbilden. Das kann durch einen Effekt zustande kommen, den man als Akquieszenz der Befragten bezeichnet. Gemeint ist, dass Personen tendenziell einer Aussage eher zustimmen als sie abzulehnen, und zwar unabhängig vom Inhalt der Aussage.

Zum anderen kann es der Fall sein, dass negativ gepolte Items eine höhere sprachliche Schwierigkeiten darstellen und mit einem solchen Item auch gleichzeitig die sprachliche Intelligenz der Proband/innen erfasst wird. Explizit angesprochen wird auch (Bühner, 2011), dass im Zusammenhang mit doppelter Verneinung die Stichprobe von entscheidender Bedeutung ist. So können Personen mit einem hohen Bildungsgrad sprachlich und strukturell komplexere Aufgaben besser handhaben. Insbesondere bei der getesteten Population, sind Schwierigkeiten mit dem Textverständnis geradezu vorprogrammiert.

Der Fragebogen soll aber explizit auch für sprachlich und kognitiv schwächere Populationen einsetzbar sein. Daher wurde in weiterer Folge auf negativ gepolte Items verzichtet, falls sich diese nur mit doppelten Verneinungen formulieren ließen.

6.4.5. Extraktion der 23 besten Items aus dem IMI

Auf Basis der Erkenntnisse zu den oben berichteten sprachlichen Schwierigkeiten wurden die ursprünglichen Items nach mathematisch-statistischen Gesichtspunkten analysiert. Inhaltliche und mathematische Argumente führten in weiterer Folge zu einem Satz bestmöglich geeigneter Items.

Eine explorative Faktorenenalyse aller 40 Items aus den sechs verwendeten Subskalen des IMI wurde mit den Daten von N = 238 Schüler/innen durchgeführt. Gemäß dem Richtwert von Freiburg (2009) für die Stichprobengröße, dass diese größer als die fünffache Zahl der Items sein soll, ist das getestete Sample ausreichend groß um interpretierbare Ergebnisse zu erhalten. Darüber hinaus ergab der Kaiser-Meyer-Olkin Test auf Eignung des Samples einen Wert von 0,822, was gut ist.

Die einzelnen Items wurden zunächst aufgrund mathematischer Argumente einer genaueren inhaltlichen und sprachlichen Begutachtung unterzogen. Auf Basis dessen wurden sie entweder aus dem Fragebogen entfernt oder beibehalten. Die mathematischen Argumente bestanden zunächst in einer Betrachtung der Korrelationen der Items zueinander, sowohl innerhalb der Subskalen, als auch innerhalb des gesamten IMI. Darüber hinaus wurden die Faktorladungen der Items auf die gewünschten Subskalen analysiert.

Nachdem sich die sechs ursprünglichen Skalen des IMI nicht reproduzieren ließen, wurde das Kaiser-Gutman-Kriterium angewendet, wonach der Eigenwert jedes Faktors größer als 1 sein soll, damit der Faktor insgesamt mehr als die Varianz einer einzelnen Variablen aufklären kann. Diesem Kriterium entsprechend wurden 5 Komponenten (Subskalen) identifiziert. Dies spiegelt sich auch im Scree-Plot wieder, der ab dem fünften Faktor flacher wird.

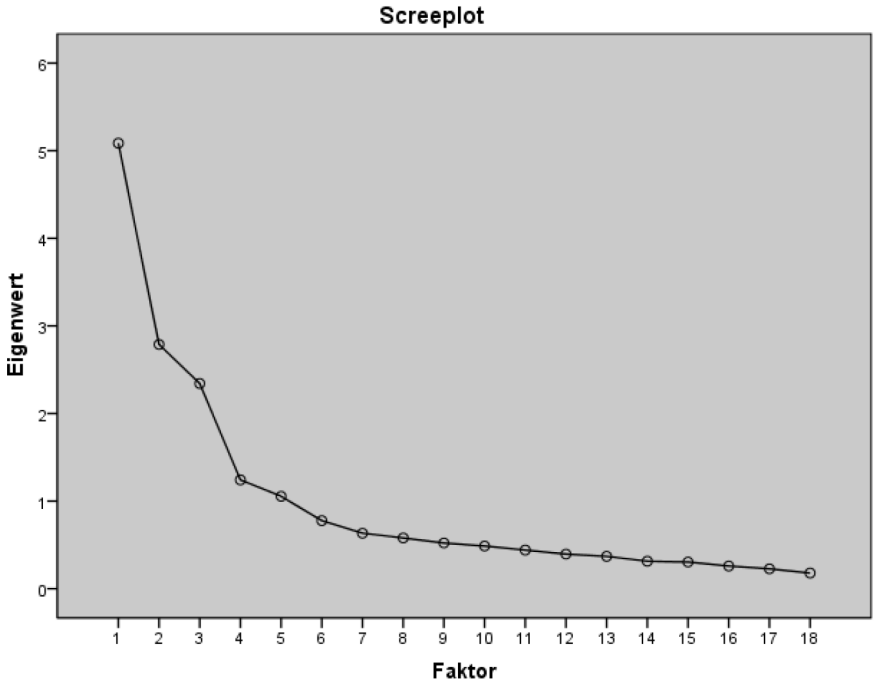


Abbildung 6.2: Scree-Plot zur Darstellung der Eigenwerte des jeweiligen Faktors. Eine Abflachung nach dem 5. Faktor ist erkennbar.

Entsprechend dieser fünf Komponenten ergab sich bei einer Extraktion mittels Hauptkomponentenanalyse bei der Varimaxrotation die folgende Faktorstruktur des IMI (Tabelle 6.4).

Dieses Modell spiegelt die bestmögliche Variante wider, was inhaltliche wie auch mathematische Argumente betrifft. Die mit diesem Modell erklärte Varianz beträgt 69,52 %. Die Reliabilität dieser Items ist mit einem Cronbach Alpha von 0,814 als hoch zu beurteilen.

Dass sich hier schlussendlich nur fünf der ursprünglich sechs Skalen des IMI abbilden ließen, kann plausibel erklärt werden. Es sind Items aus den Skalen *interest/enjoyment* und *value/usefulness*, die eine gemeinsame Skala bilden (erste Komponente). Nach Krapp (2002) ist Autonomieerleben eng damit verbunden, was einem selbst wichtig ist. Stimmen diese inneren Werte mit den Anforderungen der Aufgabenstellung überein, wird Autonomie erlebt. Dieser Argumentation zufolge sind Aufgabenstellungen, die Interesse wecken, gleichzeitig auch jene, die als wertvoll und nützlich im Rahmen der Bearbeitung erachtet werden. Insofern ist eine Verschmelzung der beiden Skalen des IMI im getesteten Sample, sowie im getesteten Kontext durchaus argumentierbar.

Rotierte Komponentenmatrix[a]

	Komponente				
	1	2	3	4	5
int5	,815				
val4	,812				
int2	,802				
int1	,760				
int7	,720				
rel3		,905			
rel4		,881			
rel8		,856			
rel7		,785			
pch3			,848		
pch4			,807		
pch5			,803		,255
pch2			,611		
pco5				,806	
pco2	,258			,772	
pco4	,409			,676	
eff2					,860
eff5					,819

Tabelle 6.4: Faktorstruktur des IMI (Werte < 0,25 wurden unterdrückt). Rotation ist in 5 Iterationen konvergiert.

Die auftretende Querladung von 0,255 bei Item pch 5 (daher beim 5. Item aus der *perceived choice* Skala) ist nicht weiter störend, da dasselbe Item auf die erwünschte Skala mit 0,803 lädt. Sind die Querladungen kleiner als die Hälfte der erwünschten Ladung, ist dies im Sinne einer eindeutigen Skalenstruktur durchaus vertretbar (Bühner, 2011). Hingegen ist die auftretende Querladung von Item pco 4 (Item 4 aus der Skala *perceived competence*) mit mehr als 60 % der Ladung auf die eigentlich gewünschte Skala problematisch.

Insgesamt ergibt sich auf Basis der obigen Analyse ein Bild des IMI, das die ursprüngliche Skalenstruktur erahnen lässt. Anzumerken ist, dass, entsprechend den Anforderungen an ein qualitativ hochwertiges und reliables Fragebogeninstrument, einzelne Skalen aus zu wenigen Items bestehen. Die *perceived competence* Skala und die *effort/impotance*-Skala beinhalten lediglich je zwei Items, wobei bei der *perceived competence* Skala eines dieser beiden Items aufgrund der hohen Querladungen nicht einmal eindeutig dieser Skala zuzuordnen ist.

Man kann nun entscheiden, wie man die weitere Vorgangsweise wählen möchte. Das ist anhängig davon, welche Anforderungen man an ein Messinstrument stellt. Soll es psychometrisch exakt und hochreliabel sein, oder genügt es, ein leidlich gut funktionierendes Instrument zu haben? Um das zu entscheiden ist die Frage nach dem Stellenwert zu beantworten, den das Messinstrument bzw. die damit generierten Daten, im gesamten Untersuchungsdesign einnehmen. Damit verbunden ist, die Frage zu entscheiden, wie viel Testzeit zur Verfügung stehen soll und den Schüler/innen zumutbar ist.

Basierend auf Überlegungen bezüglich Sprachverständlichkeit und Polarität wurden in Kombination mit inhaltlichen Kriterien (Zugehörigkeit einzelner Items zu bestimmten Skalen, Schülerinterviews zur Sprachverständlichkeit der Items) und den oben berichteten mathematischen Kriterien (explorative Faktorenanalysen und Korrelationen) jene Items extrahiert, die am ehesten geeignetsten erschienen, das Konstrukt der Motivation entsprechend der SDT abzubilden.

Dabei wurden gegenüber der ersten Übersetzung (vgl. Anhang 13.2) folgende sprachliche Adaptierungen vorgenommen:

- Die Items eff 2, eff 5, pch3 und pch 4 wurden so umformuliert, dass die negative Polung gegenüber der ersten Übersetzung wegfiel.
- Das Item val 4 wurde gekürzt.
- Die Items pch 1, int 7 und rel 8 wurden in freierer Übersetzung sprachlich adaptiert.

Es ergab sich ein Fragebogen, der nunmehr aus 23 Items bestand (vgl. Anhang 13.3). Dieser Fragebogen wurde auch parallel zur Datensammlung für die Wissenstests bezüglich Elektrizitätslehre eingesetzt und zwar nach dem Tutoring.

Einer der größten Nachteile bestand darin, dass nicht alle Skalen mit einer als ausreichend erscheinenden Anzahl an Items abgebildet werden konnten. Der Wunsch, ein Instrument mit einer genügenden Anzahl an Items zu entwickeln, hatte aber zunächst den Ausschlag dafür gegeben, bereits bestehende Instrumente nicht zu übernehmen. Daher musste weitere Entwicklungsarbeit in das Instrument gesteckt werden.

6.4.6. Ergebnis der zweiten Pilotierung mit den besten 23 Items aus dem IMI

Diese verbesserte, sprachlich adaptierte und gekürzte Version des IMI, die nunmehr aus 23 Items bestand, wurde an N = 138 Schüler/innen aus sechs Klassen der Hauptschule und der AHS erprobt. Entsprechend dem angegebenen Richtwert, dass die Anzahl der Versuchspersonen größer als die etwa fünffache Itemanzahl sein soll, ist die Stichprobengröße wiederum ausreichend.

Explorative Faktorenanalysen in diesem neuen Sample ergaben jedoch, dass die in der ersten Pilotierung erhaltene Faktorstruktur nicht reproduzierbar war, auch wenn man sich auf die in Tabelle 6.4 beschriebenen Items beschränkt. Lediglich die Skalen *interest* und *perceived competence* blieben auch in dieser Stichprobe stabil.

Alle anderen Skalen zeigten zudem eine weitere unerwünschte Stichprobenabhängigkeit, die von der Klasse abhing und die am folgenden Beispiel gezeigt werden soll:

Analysiert man im vorliegenden Sample klassenweise, so ergeben sich für das Motivationsinstrument völlig unterschiedliche Faktorstrukturen. Tabelle 6.5 und Tabelle 6.6 stellen die Ergebnisse der Faktorenanalysen in einer Hauptschulklasse und in einer AHS-Klasse dar, also zweier Klassen, die unterschiedlichen Schultypen angehören und von denen man annehmen kann, dass daher das kognitive Niveau unterschiedlich war.

Rotierte Komponentenmatrix[a,b,c]

	Komponente				
	1	2	3	4	5
mot10	,876				
mot17	,846				
mot8	,810	,342			
mot3	,775	,426			
mot5	,736				,356
mot6	,653	,444			,316
mot2	,621	,307			
mot19		,808			
mot14		,798			
mot20		,673			
mot1	,483	,649	,308		
mot9		,604			,360
mot16		,596	,383	,407	
mot22	,490	,526	,461		
mot18			,803		
mot12	,375		,760		
mot21			,738		
mot23				,821	
mot11				,761	
mot15	,453			,531	
mot7		,358			,702
mot13			,557		,683
mot4	,301			,409	,625

Tabelle 6.5: Faktorenstruktur von 23 Items des IMI für Klasse 2 (Hauptschule) nach einer Varimax-Rotation mit Hauptkomponentenanalyse (Konvergenz in 8 Iterationen).

Rotierte Komponentenmatrix[a,b,c]

	Komponente				
	1	2	3	4	5
mot17	,800				
mot12	,785				
mot5	,768		,329		
mot8	,684				
mot14	,606		,401		
mot22	,601				
mot9	,433	,427		,418	
mot19	,348	,802			
mot16	,432	,754			
mot1		,753			
mot6		,751			
mot20		,632			,439
mot21			,791		
mot18			,740		
mot13			,731		
mot15		,468	,495	,346	
mot2				,773	
mot10		,452		,692	
mot3	,385	,309		,620	
mot4		,358	,414	-,419	
mot23					,870
mot11					,838
mot7			,363		,497

Tabelle 6.6: Faktorenstruktur von 23 Items des IMI für Klasse 3 (AHS) nach einer Varimax-Rotation mit Hauptkomponentenanalyse (Konvergenz in 6 Iterationen).

Eine derartige Stichprobenabhängigkeit ist für ein Messinstrument, das in der gesamten Sekundarstufe 1, unabhängig vom Schultyp, eingesetzt werden soll, unerwünscht, da es keine reliablen Ergebnisse liefern kann.

Nach dieser zweiten Pilotierung stellt sich die Datenlage somit widersprüchlich zu den Ergebnissen der ersten Pilotierung dar. Daher wurde in weiterer Folge eine gänzlich andere Vorgangsweise gewählt, als nach der ersten Pilotierung indiziert erschien. Statt das IMI zu verbessern, wurde begonnen, Items konstruktorientiert neu zu formulieren. Diese Konstruktionsarbeit stellt jedoch einen erheblichen Aufwand dar, soll schlussendlich doch ein mehrdimensionales Messinstrument das Ergebnis sein, das üblichen Anforderungen die Reliabilität und Validität betreffend genügt.

6.5. Konstruktion einer Skala (Effort/Importance)

Aus Gründen der Stichprobenabhängigkeit des in Kapitel 6.4 untersuchten Messinstrumentes und der wenigen passenden Items zur *effort/importance*-Skala wurden Überlegungen angestellt, zumindest für die *effort/importance*-Skala von Grund auf neue Items zu konstruieren. Es wurde somit klar, dass es im Zuge dieser Studie kein Messinstrument zur Motivation geben konnte, das allein auf dem IMI basierte.

In der deutschen Übersetzung für *effort/importance*, wurde die Wortwahl Einsatzbereitschaft/Wichtigkeit für die Arbeit mit den Schüler/innen gewählt.

Die für unsere Zwecke wichtigste Skala scheint die *effort/importance*-Skala zu sein, die Einsatzbereitschaft und Wichtigkeit misst und somit als Maß dienen soll, wie stark sich die Tutor/innen im CAPT-Projekt engagieren. Dieses Engagement wiederum soll mit den Ergebnissen der Wissenstests nach der CAPT Intervention verglichen werden, um mögliche Zusammenhänge identifizieren zu können. Deshalb war die *effort/importance*-Skala Ausgangspunkt der Entwicklungsarbeit neuer Items. Für der Konstruktion neuer Items wurde der *Four Building Blocks Approach* von Wilson (2005) gewählt.

6.5.1. Grundlagen der Testtheorie und Basisideen von Messungen

Wie ein großer Teil der bestehenden psychometrischen Testverfahren sollen die hier vorgestellten, neu entwickelten Items der klassischen Testtheorie genügen. Diese wird auch als *true score theory* bezeichnet und kann auch als eine Theorie der Messfehler bezeichnet werden. Die Grundannahme besteht darin, dass die variierenden Messwerte, die bei wiederholten Messungen für eine Person erhoben werden, auf systematische Einflüsse (z.B. Training, Instruktion) und unsystematische Messfehler zurückzuführen sind.

Betrachtet man den *un*systematischen Messfehler, so ist er bei wiederholter Messung an einer Versuchsperson häufig normalverteilt (intraindividuelle Verteilung). Die Ursachen dafür können auf innere Einflüsse (z.B. Tagesverfassung) oder äußere Einflüsse zurückzuführen sein (z.B. Wetterlage zum Messzeitpunkt). Jedenfalls wird der Messfehler als zufällig betrachtet.

Die Basis der klassischen Testtheorie bilden zwei Grundannahmen:

- Es gibt einen wahren Wert für eine Person. Dieser stellte den Erwartungswert der beobachteten Messwerte x_i für eine Person dar. Der Erwartungswert ist ein Schätzer für den Mittelwert aus vielen, wiederholten Messungen an einer Person. Außerdem soll der beobachtete Wert auch etwas tatsächlich Existierendes (per-fiat-Messung[14]) messen.
- Der Messfehler (zu einem Zeitpunkt) für eine Person ist die Differenz aus dem beobachteten Wert und dem wahren Wert für diese Person.

Aus diesen Grundannahmen können zahlreiche Folgerungen abgeleitet werden.

- Die Varianz der beobachteten Messwerte ist ein Schätzer für die Varianz des Messfehlers $\sigma^2(E)$.
- Der beobachtete Wert (X) für eine Person setzt sich immer aus dem wahren Wert (T) und einem Messfehler (E) zusammen: X = T + E.
- Der Erwartungswert des Messfehlers ist somit für unendlich viele Messungen Null (Bühner, 2011). Er mittelt sich daher aus.

[14] Aus dem Lateinischen übersetzt „so sei es“.

- Aus diesem Grund ist der Mittelwert der beobachteten Werte ein Schätzer für den Erwartungswert in der Population.

Die erste Grundannahme ist sehr umfangreich. Sie bedeutet nämlich, dass die gesammelten Daten korrekt sind, also die Befragten die Wahrheit sagen. Außerdem ist hier implizit als Basisannahme von Messungen enthalten, dass die gesammelten Daten mit den wahren Werten des Merkmals korrespondieren. Eine Variation im Merkmal des zu untersuchenden Objektes verursacht demnach eine Variation in der Observablen.

Dieser Zusammenhang ist in Abbildung 6.3 dargestellt. Als Beispiel wurde die Eigenschaft „Temperatur“ gewählt. Eine bestimmte Temperatur führt dazu, dass das Thermometer etwas anzeigt. Liest man das Thermometer ab, kann man der angezeigten Zahl die Bedeutung der Temperatur zuordnen. In diesem Fall ist der Zusammenhang streng deterministisch und der Rückschluss von der abgelesenen Zahl auf die Temperatur kann fehlerfrei erfolgen. Das bedeutet allerding noch nicht, dass auch die Ablesung der Zahl am Thermometer fehlerfrei vor sich ging.

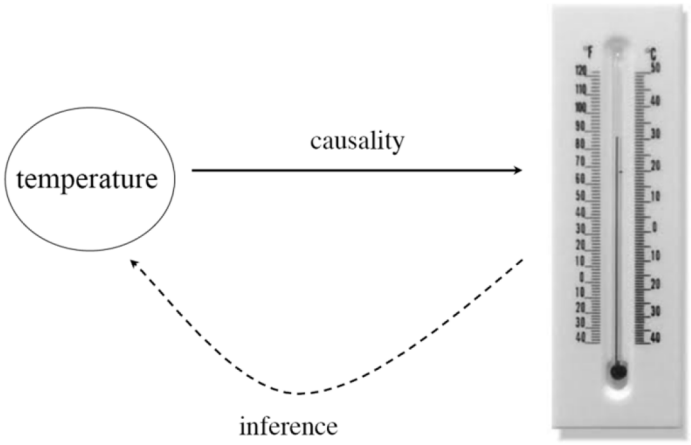


Abbildung 6.3: Grundidee jeder Messung: Ein Merkmal führt zu einer beobachtbaren Veränderung, von der aus wiederum Rückschlüsse auf das Merkmal möglich sind. Mit freundlicher Genehmigung von A. E. Maul, University of Colorado

Im Falle latenter Variabler, wie es z.B. bei vielen psychometrischen Größen der Fall ist, ist der Zusammenhang zwischen dem Merkmal und der Observablen nicht streng deterministisch, sondern hat Wahrscheinlichkeitscharakter (Abbildung 6.4).

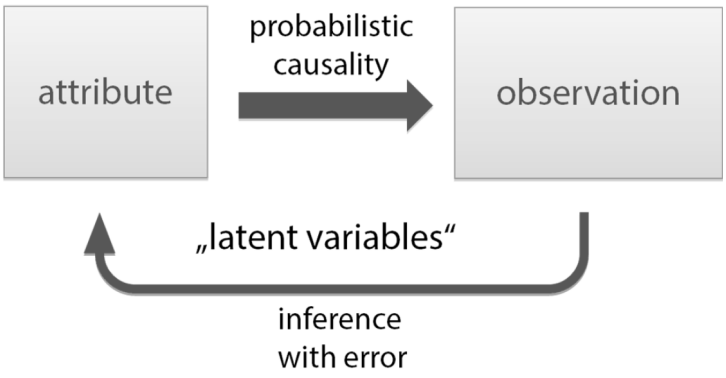


Abbildung 6.4: Zusammenhang zwischen Eigenschaft und Beobachtung im Falle latenter Eigenschaften

Der Rückschluss von der Beobachtung auf das Merkmal ist daher bereits fehlerbehaftet, unabhängig davon, wie genau die Beobachtung stattgefunden hat.

Im Falle der hier zu entwickelnden Items handelt es sich für das Personenmerkmal „Anstrengungsbereitschaft“ um eine latente Variable und daher um eine prinzipiell fehlerbehaftete Inferenz. Diesem Umstand wird in der Anwendung der passenden statistischen Prozeduren und der Einführung einer Irrtumswahrscheinlichkeit α bei den Hypothesentests Rechnung getragen.

Der *Four Building Block Approach* (Wilson, 2005) schlägt nun vier Schritte (*building blocks*) vor, mit denen eine Konstrukt abgebildet werden kann (Abbildung 6.5). Darüber hinaus beinhaltet diese Vorgangsweise auch bereits wichtige Überlegungen zur Reliabilität der Items.

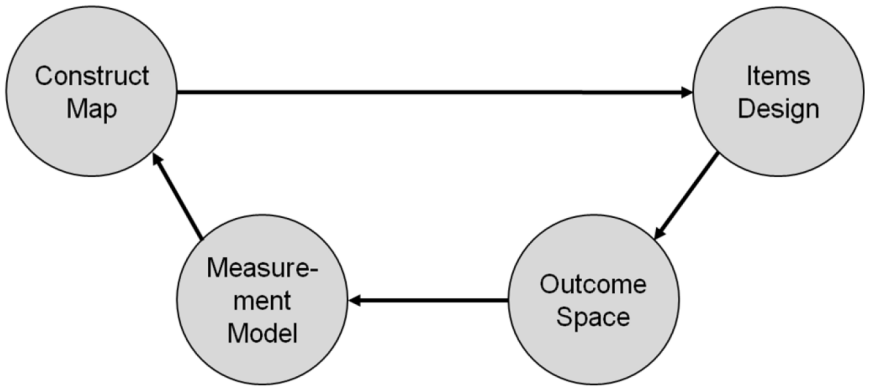


Abbildung 6.5: Die Grundelemente des *Four Building Blocks Approach* zur Modellierung eines Konstrukts. Nach (Wilson, 2005, S. 17).

Im Folgenden werden nun die vier Blöcke zur Konstruktion geeigneter *effort/importance*-Items abgearbeitet.

6.5.2. Die vier Building Blocks

Block 1: Die Construct Map zur Spezifizierung des Konstrukts

Eine *Construct Map* stellt die Visualisierung eines bestimmten Konstrukts dar. Bevor diese Visualisierung konkret umgesetzt wurde, wurden Überlegungen angestellt, auf welcher Basis „Einsatzbereitschaft/Wichtigkeit“ überhaupt abgebildet werden können. Es erschien nötig Evidenzen zu finden, die später einen Rückschluss zur Anstrengungsbereitschaft möglich machen.

Um die eigentliche *Construct Map* (vgl. Tabelle 6.7) zu erhalten, wird die Dimension des Konstrukts kontinuierlich in Richtung ihrer (extremen) Ausprägungen zerlegt. Dem zugrunde liegt die Annahme, dass es überhaupt eine qualitative Ordnung des abzubildenden Merkmals gibt, die dem Konstrukt inhärent ist. Es ist sinnvoll, sich zunächst die extremen Merkmalsausprägungen vor Augen zu führen und diese zu beschreiben. Unter der Annahme, dass man das Konstrukt auf einem Kontinuum abbilden kann, können danach die dazwischen liegenden Zustände beschrieben werden.

In diesem Fall sind die beiden extremen Merkmalsausprägungen durch die Formulierungen „Schüler/in steckt wenig Einsatzbereitschaft ins CAPT“ und „Schüler/in steckt viel Einsatzbereitschaft ins CAPT“ repräsentiert. Auf der linken Seite der *Construct Map* werden zwischen diesen Extremen die unterschiedlichen Ausprägungen des Merkmals „Einsatzbereitschaft“ beschrieben, die die Versuchspersonen zeigen können. Im speziellen Fall sind das die Schüler/innen, genauer die Tutor/innen, die sich mehr oder weniger im Tutoring Prozess engagieren. Auf der rechten Seite wird versucht, zu den Merkmalausprägungen korrespondierende Evidenzen festzumachen. Tabelle 6.7 zeigt einen ersten Versuch für diese Evidenzen.

Schüler/innen stecken viel Einsatzbereitschaft ins CAPT	
Schüler/innen	**Antworten zu den Items**
Schüler/innen engagieren sich beim CAPT	strengen sich beim CAPT sehr an; bereiten sich auf CAPT genau vor
Schüler/innen verhalten sich neutral gegenüber CAPT	bereiten sich dafür vor, weil sie dazu angehalten werden
Schüler/innen lehnen CAPT ab	bereiten sich selbst gar nicht auf CAPT vor; sind froh, wenn es vorüber ist
Schüler/innen stecken wenig Einsatzbereitschaft ins CAPT	

Tabelle 6.7: *Construct Map* zum Konstrukt Einsatzbereitschaft/Wichtigkeit (*effort/importance*)

Dieser erste Versuch führte nicht zu zufrieden stellenden Ergebnissen, da der Prozess der Klärung, was das Konstrukt denn eigentliche bedeutet, zuvor nicht vollständig abgeschlossen gewesen ist. So ist „engagieren“ nichts, das sich auf eine Evidenz zurückführen lässt. Für eine evidenzbasierte Variablenklärung wurden in weiterer Folge die eigentlichen Expert/innen herangezogen: die Schüler/innen selbst. Sie sollten zur Klärung beitragen, was Einsatzbereitschaft und Wichtigkeit einer Sache für sie selbst bedeuten und an welchen Beobachtungen man diese Eigenschaften erkennen kann.

Nach mehreren Versuchen, die Fragestellung an die Schüler/innen zu präzisieren, wurde von Seiten der Forscher/innen das Konstrukt auf die Frage herunter gebrochen: „Woran *sieht* man, dass sich jemand angestrengt?“. Die Fragestellung zielte somit auf die Sammlung von Evidenzen ab, die mit dem Konstrukt verbunden sind. In zahlreichen Rückmeldungen und Klärungen der Fragstellung wurde immer wieder darauf hingewiesen, dass es sich hier um die Frage nach *Beobachtbarem* handelt und nicht um Einschätzungen.

Aus dieser Befragung von insgesamt 51 Jugendlichen ergab sich die Liste folgender Evidenzen, die im Einzelnen besser oder schlechter beobachtbar waren und sich zur Formulierung von Items verwenden ließen. Antworten, die sich auf körperliche bzw. sportliche Anstrengungen bezogen (z.B. „Person schwitzt stark“) wurden nicht in die Liste aufgenommen.

- Schüler/innen aus der 11. Schulstufe AHS, N=5
  - Wie viel Zeit?
  - Wie viele Quellen?
  - Wie intensiv?
  - etwas anderes dafür nicht gemacht
  - ich würde diese Arbeit nicht gegen etwas anderes tauschen
- Schüler/innen aus der 8. Schulstufe HS, N=20
  - Person macht mit
  - Person macht etwas sofort
  - Person fragt nach
  - Person will gleich ausprobieren
- Zwei Kinder, 14 und 17 Jahre alt
  - Person verzichtet auf etwas anderes
  - Person fragt nach
- Schüler/innen aus der 8. Schulstufe AHS, N=24
  - wenn Person nicht locker lässt, bis sie auf ein gutes Ergebnis gekommen ist
  - Versuch offenen Fragen auf den Grund zu gehen
  - Person fragt nach ||
  - Person arbeitet oft und lange
  - Person setzt ehrgeiziges Ziel
  - Person übernimmt schwierige Arbeiten
  - Wenn Person nicht dauernd sagt: „Wann kann ich endlich aufhören?“
  - man muss Person zwingen aufzuhören
  - Person gibt nicht auf, wenn es nicht klappt
  - Zeitaufwand |||
  - Person denkt statt: “Na endlich geschafft“ eher „Das war cool“
  - Person sucht Beratung bei anderen
  - Person fragt Fachleute um Meinung
  - Person macht viel (und sitzt nicht nur herum)
  - Person macht keine/kaum Pausen ||

– Person arbeitet schnell/zügig
– Person ist fröhlich, wenn sie fertig ist

Eine genauere Analyse dieser Schülereinschätzungen zeigt, dass sich zwei Kategorien extrahieren lassen, die als Evidenzen für Einsatzbereitschaft und Wichtigkeit gelten: ein Zeitfaktor und ein Faktor, den man als Qualität der Arbeit bezeichnen kann. Indem die Aussagen zu den einzelnen Kategorien nach Merkmalsausprägung geordnet wurden, wurden zu diesen beiden Bereichen neue *Construct Maps* entwickelt.

Die Vorgangsweise, das Konstrukt in zwei Bereiche von Evidenzen zu zerlegen ist nicht ganz ungefährlich. Einerseits ist zwar die Möglichkeit gegeben, dass dadurch mehrere Aspekte des Konstrukts *Einsatzbereitschaft/Wichtigkeit* abgebildet werden. Das bedeutet nämlich, dass die Gefahr der Unterrepräsentation des Konstrukts durch die vielfältigen Evidenzen minimiert wird, dass die Evidenzen eher dazu geeignet sind das gesamte Konstrukt abbilden und nicht nur Teilaspekte davon. Dadurch steigt die Validität. Auf der anderen Seite kann es jedoch passieren, dass das Konstrukt bei konfirmatorischen Faktorenanalysen in zwei Teilkonstrukte zerfällt, die keine gemeinsame Skala ergeben.

Schüler/innen betreiben einen hohen Zeitaufwand beim CAPT

Schüler/innen	**Antworten zu den Items**
Schüler/innen haben viel Freizeit/Pause investiert; wollen es sehr gut machen	haben viel Zeit investiert; Quellen studiert; haben auf etwas Pause/Freizeit verzichtet
Schüler/innen haben etwas Freizeit/Pause investiert; wollen es gut machen	strengen sich beim CAPT an; bereiten sich auf CAPT vor
Schüler/innen haben nur das Notwendigste gemacht	haben etwas Zeit investiert, weil sie dazu angehalten werden
Schüler/innen haben gewartet, bis es vorbei ist; wollten unentdeckt bleiben	bereiten sich selbst gar nicht auf CAPT vor; sind froh, wenn es vorüber ist

Schüler/innen betreiben einen geringen Zeitaufwand beim CAPT

Tabelle 6.8: *Construct Map*, die den Zeitfaktor als Maß für Engagement im Tutoring Prozess widerspiegeln soll.

Tabelle 6.8 zeigt die *Construct Map* für den Zeitfaktor. Damit ist die Zeit gemeint, die die Schüler/innen in ihre Aufgabe investiert haben. Der Unterschied dieser *Construct Map*

im Vergleich zu der aus Tabelle 6.7 liegt vor allem darin, dass „sich engagieren“ nunmehr auf konkrete Evidenzen zurückgeführt werden kann. Außerdem ist die Unterteilung in der Merkmalsausprägung feiner.

Tabelle 6.9 stellt eine *Construct Map* dar, die die Qualität der Arbeit der Schüler/innen beim CAPT erfassen soll.

Schüler/innen liefern qualitätsvolle Arbeit beim CAPT

Schüler/innen	**Antworten zu den Items**
Schüler/innen sind mit ihrer Arbeit selbst zufrieden	lassen nicht locker, bis sie auf ein gutes Ergebnis kommen
Schüler/innen fragen nach und klären dann offene Fragen	gibt nicht auf, wenn es nicht klappt
Schüler/innen erledigen alles gründlich und schnell	arbeiten schnell und zügig
Schüler/innen gehen die Themen beim CAPT schnell durch	denken: „endlich geschafft“

Schüler/innen schludern[15] die Arbeit beim CAPT hin

Tabelle 6.9: *Construct Map*, die die Qualität der Arbeit im Tutoring Prozess widerspiegeln soll.

Auf Basis dieses Prozesses folgte der zweite Schritt, nämlich die Konstruktion konkreter Items auf Basis der *Construct Maps*, die nunmehr an Evidenzen orientiert waren.

Block 2: Item Design

Auf Basis der Construct Maps und der Befragungen der Schüler/innen wurden die folgenden Items konstruiert. Sie teilen sich in zwei thematische Blöcke, den „Zeitfaktor“ und „Qualität der Arbeit“ entsprechend der *Construct Maps*.

- *Items zu „Zeitfaktor“*

  Ich habe mich sofort auf meine Aufgaben gestürzt.

  Ich wollte diese Aufgaben gleich mit dem anderen Kind ausprobieren.

  Ich wollte sofort mit dieser Aufgabe anfangen.

  Ich habe mir nie gedacht: „Wann kann ich endlich aufhören?“

[15] ohne viel Sorgfalt arbeiten, nachlässig arbeiten

Am liebsten hätte ich noch nicht aufgehört, daran weiter zu arbeiten.

Man musste mich direkt zwingen aufzuhören.

Ich habe viel Zeit in diese Aufgabe gesteckt.

Ich habe bei dieser Aufgabe viel gearbeitet und bin nicht nur herumgesessen.

Ich habe bei dieser Aufgabe kaum Pausen gemacht.

Ich habe bei dieser Aufgabe kaum an etwas anderes gedacht.

Ich habe zügig gearbeitet.

Ich habe sogar auf etwas Freizeit (Pause) verzichtet, um bei dieser Aufgabe gut zu sein.

Ich habe viel Zeit für die Vorbereitung investiert / aufgewendet.

Ich habe lange an dieser Aufgabe gearbeitet.

Ich habe bis zum Schluss der Arbeitszeit an dieser Aufgabe gearbeitet.

- *Items zu „Qualität der Arbeit“*

  Ich habe versucht, offenen Fragen auf den Grund zu gehen.

  Ich habe weiter gearbeitet, bis ich mit dem Ergebnis zufrieden war.

  Es ist mir wichtig, dass ich mit meinem Ergebnis zufrieden bin.

  Ich habe nachgefragt, bis der andere / die andere wirklich alles verstanden hat.

  Ich habe mir selbst ein (ehrgeiziges) Ziel gesetzt.

  Ich würde diese Arbeit nicht gerne gegen etwas anderes tauschen.

  Ich übernehme auch schwierige Aufgaben.

  Ich habe mir gedacht: „Das war cool.“

  Ich habe nicht aufgegeben, auch wenn es nicht gleich geklappt hat.

  Ich habe mich fröhlich gefühlt, als ich fertig war.

Aus der Liste dieser Vorschläge wurden die grau unterlegten Items schließlich ausgewählt, da sie das Konstrukt hinsichtlich seiner Breite am besten repräsentieren. Diese Einschätzung wurde im Gespräch mit Expert/innen im Rahmen von Arbeitsgruppentreffen getroffen.

Alle diese Items haben gemeinsam, dass sie in der *Construct Map* sehr hoch oben angesetzt sind. Sie alle repräsentieren das Konstrukt Einsatzbereitschaft/Wichtigkeit daher in einem hohen Maße. Das sollte per se kein Nachteil sein, da die Items nicht

durch eine dichotome Skala abgefragt werden, sondern zu Gunsten einer Likertskala entschieden wurde, wie im nächsten Abschnitt ausführlicher argumentiert wird. Dennoch ist zu bedenken, dass der Mittelwert nicht aussagekräftig interpretierbar ist und lediglich Vergleiche zulässt.

Block 3 und Block 4: Der *Outcome Space* und das *Measurement Model*

Will man vom *Outcome Space* Rückschlüsse auf das zugrunde liegende Konstrukt machen, bedarf es einer Messvorschrift (*Measurement Model*). Hierzu gibt es im Wesentlichen zwei mögliche Ansätze: Der eine fokussiert auf die Items und ihre innere Konsistenz. Dabei wird einer Testperson eine bestimmte Anzahl von Items vorgelegt, die alle auf dasselbe Konstrukt abzielen. Deren Scores werden dann addiert (oder es wird der Mittelwert gebildet). Ein hoher Gesamtscore lässt auf eine hohe Merkmalsausprägung hinsichtlich des Konstrukts schließen. Dieser Ansatz stellt daher eine Verbindung zwischen Gesamtscore und Konstrukt her, obwohl nicht immer klar formuliert wird, wie genau diese Inferenz zustande kommt. Für die klassische Testtheorie stellt dieser Ansatz die Basis dar.

Der andere Ansatz fokussiert auf die Items und gibt Aussagen vor, denen die Testpersonen zustimmen oder sie ablehnen können. Der Gesamtscore richtet sich danach, wie vielen Aussagen die Testperson zustimmen konnte.

Von der Wahl zwischen den beiden Zugängen hängt bereits die Modellierung des *Outcome Space* ab.

Der zweite Zugang wurde vor allem von Guttman in den 1940er-Jahren formalisiert (Wilson, 2005). Statt wie bei Likert-skalierten Items Antwortmöglichkeiten anzugeben von „Ich stimme sehr zu“ bis zu „Ich stimme gar nicht zu“ werden diese Kategorien durch bestimmte Aussagen ersetzt. Die Probanden können diesen Aussagen zustimmen oder auch nicht. Üblicher Weise werden diese Aussagen immer extremer formuliert, sodass eine Zustimmung zur n-ten Aussage bedeutet, dass die Testperson auch allen davor zugestimmt hat, weil sie weniger streng waren, bzw. das Konstrukt in schwächerer Weise repräsentiert haben. Diese Möglichkeit, Items zu gestalten hat den Vorteil, dass klar ausformuliert ist, welcher Aussage man zustimmt. Bei der Formulierung „Ich stimme stark zu“ bleibt immer offen, was die einzelne Testperson unter „stark zustimmen“ versteht. Gleiche Angaben auf einer Likert-Skala können demnach sehr unterschiedliche

Ausprägungen des Konstrukts bedeuten. Das ist *ein* Grund dafür, dass Testergebnisse mit Likert-skalierten Items stark vom Kontext und dem kulturellen Hintergrund der getesteten Personen abhängen. Darüber hinaus kann es vorkommen, dass ein Proband einer Sache mit großer Überzeugung neutral gegenüber steht. Diese Merkmalsausprägung können Likert-skalierte Items nicht abbilden.

Dennoch wurde im Rahmen dieser Arbeit darauf verzichtet, Guttman-skalierte Items zu entwickeln. Einerseits, weil der Kontext (CAPT) und der kulturelle Hintergrund des Samples (Schüler/innen der Sekundarstufe 1 im Großraum Wien) sehr homogen waren. Andererseits, weil die Entwicklung valider Distraktoren für diese Art von Items eine Entwicklung und Erprobung in mehreren Zyklen bedeutet, was den Rahmen dieser Arbeit sprengt. Schlecht entwickelte Distraktoren stellen gegenüber Likert-skalierten Items sogar einen Nachteil dar. Darüber hinaus sollte die Entwicklung des Motivationsinstruments so rasch von statten gehen, dass es noch im Rahmen von CAPT verwendet werden kann. Die Beforschung der Motivation stellte im hier beschriebenen Projekt nämlich lediglich einen Teilaspekt dar. Es ist auch zu bedenken, dass Guttman-skalierte Items den Nachteil haben, dass viel Text zu lesen ist, was wiederum den zeitlichen Aufwand beim Testen erhöht und letztlich auch das Textverständnis der Schüler/innen testet, was in diesem Zusammenhang nicht erwünscht war.

Es wurden daher Items entwickelt, deren *Outcome Space* Likertskalen bilden. Allerdings wurden bei der Formulierung der Items Anstrengungen unternommen, das Konstrukt präzise abzubilden anstatt Ad-hoc-Formulierungen zu verwenden, wie dies z.B. bei den IMI-Items der Fall ist. Eine weitgehend präzise Abbildung des Konstrukts sollte durch die Verwendung der *Construct Maps* sichergestellt worden sein.

Die Messvorschrift selbst besteht darin, dass über die einzelnen Items Mittelwerte, oder alternativ Summenscores, gebildet werden sollen. So können einerseits Klassen miteinander verglichen werden, andererseits können Vergleiche zwischen den Schüler/innen einer Gruppe durchgeführt werden.

6.5.3. Quellen der Validität des Messinstrumentes

Die Validität eines Messinstrumentes kann im Wesentlichen durch zwei Dinge eingeschränkt werden:

1) Das Konstrukt ist unterrepräsentiert. Damit ist gemeint, dass eine Variation in *Teilen* des Konstrukts zu einer Variation in der Beobachtung führt. Umgekehrt schließt man dann aus einer veränderten Beobachtung auf ein verändertes Attribut, obwohl sich tatsächlich nur Teile des Attributes verändert haben. Dieser Mangel an Validität betrifft das gesamte Konstrukt.

2) Andere Faktoren, die außerhalb des Konstrukts liegen, beeinträchtigen die Validität. Das bedeutet, dass man nicht nur das Konstrukt misst, sondern auch andere Variable, die man vielleicht nicht berücksichtigt hat.

Im Folgenden werden Überlegungen und Evidenzen für die Validität des Messinstrumentes angeführt und in Anlehnung an Wilson (2005) diskutiert.

Um der Unterrepräsentation des Konstrukts zu entgehen, ist der beschriebene Weg des *Four Building Blocks Approach* per se bereits eine gute Absicherung. Die Durchführung der vier Schritte dient bereits als Beleg einer sorgfältigen Repräsentation des Inhaltes des Konstrukts.

Als eine weitere Bestätigung der Validität kann der Antwortprozess als solcher angesehen werden. Dazu wurden mit N = 10 Schüler/innen Exitinterviews zur Verständlichkeit der Items geführt und das individuelle Antwortverhalten reflektiert.

Weitere Überlegungen basieren auf der internen Struktur des Messinstrumentes. Hierbei ist es wiederum so, dass Überlegungen zur *Construct Map* und deren innere Struktur bereits dazu führen, dass das Messinstrument das Konstrukt gut repräsentiert. Weiter führende Analysen wie die Itemanalysen (Abbildung 6.6 und Abbildung 6.7) lassen darauf schließen, dass das Itemdesign eben diese Strukturen des Konstrukts widerspiegelt.

6.5.4. Testung der neuen Items

Die neu entwickelten Effort-Items wurden gemeinsam mit den restlichen Items zu den anderen Subskalen des IMI an N = 192 Schüler/innen getestet. Wenn man davon ausgeht, dass sich die Subskalen des IMI gegenseitig nicht beeinflussen, da sie ja auf unterschiedliche Konstrukte fokussieren, so ist es durchaus zulässig, aus dieser Datenmenge die Effort-Items zu nehmen und diese separat zu analysieren. Gemeinsam mit den 5 Items, eff1 bis eff 5, aus dem IMI wurden die neu konstruierten Effort-Items

eff 6 bis eff 17 wurde eine Langversion des Motivationsfragebogens geschaffen. Diese ist dem Anhang 13.4 zu entnehmen.

Um geeignete von weniger gut geeigneten Items zu separieren, wurden zwei Vorgangsweisen gewählt: Für einen ersten Eindruck, wurden die Korrelationen der Items untereinander betrachtet und danach eine Itemanalyse durchgeführt, bei der der Score aller Testpersonen auf das Item mit dem Score auf das gesamte Konstrukt verglichen wurde. Passt ein Item gut in das Konstrukt, so scoren Testpersonen auf das Item genauso hoch wie auf das gesamte Konstrukt.

Die Analyse der Pearson Korrelationen ergab, dass die Items eff 1, eff 3, eff 4, eff 5 und eff 11 auf einem höchstsignifikanten Niveau ($p < 0,01$) hoch korreliert ($\rho > 0,5$) sind.

Die Items eff 2, eff 9, eff 10 und eff 14 korrelieren immerhin mit mittleren Korrelationen ($\rho > 0,4$) auf einen Signifikanzniveau von $p < 0,01$. Nun sind Korrelationen nicht gleichbedeutend mit Kausalitäten und daher mit Vorsicht zu interpretieren. Dennoch geben sie, gemeinsam mit den Überlegungen, die zur *Construct Map* angestellt wurden, einen Hinweis auf die Konsistenz des Konstrukts und nicht zuletzt auf seine Validität und Reliabilität. Außerdem sind die Korrelationen wichtig, um die Beiträge einzelner Items zu Cronbach's α beurteilen zu können: Denn hier verbessern auch nicht hoch korrelierte Items („randständige" Items) α, tragen aber inhaltlich nicht zur Klärung des Konstrukts bei.

Die Vorgangsweise der Itemanalyse sei an einem Beispielitem erklärt. In einem Streudiagramm wird auf der Abszisse der Score jeder Testperson auf das gesamte Konstrukt „*Effort/Importance*" aufgetragen. Auf der Ordinate wird der Score der Testperson auf das einzelne Item aufgetragen. Passen die Items gut zusammen, also entspricht der Score auf das Konstrukt dem Score auf das Item, müsste eine Trendlinie daher der ersten Mediane entsprechen. Liegt die Trendlinie darüber, so überschätzt das Item das Konstrukt. Das bedeutet, dass das Item eine stärkere Merkmalsausprägung angibt als das gesamte Konstrukt. Liegt die Trendlinie darunter, so unterschätzt das Item das Konstrukt. Besonders heikel ist es, wenn sich die beiden Linien kreuzen, denn dann werden die extremen Merkmalsausprägungen über- oder unterschätzt. Dort zeigt sich

aber gerade die Güte eines Items, da es ein Hinweis auf seine Trennschärfe ist, wenn es einen möglichst breiten Bereich an Merkmalsausprägungen konsistent darstellen kann.

Beispiel für ein das Konstrukt überschätzendes Item (Abbildung 6.6):

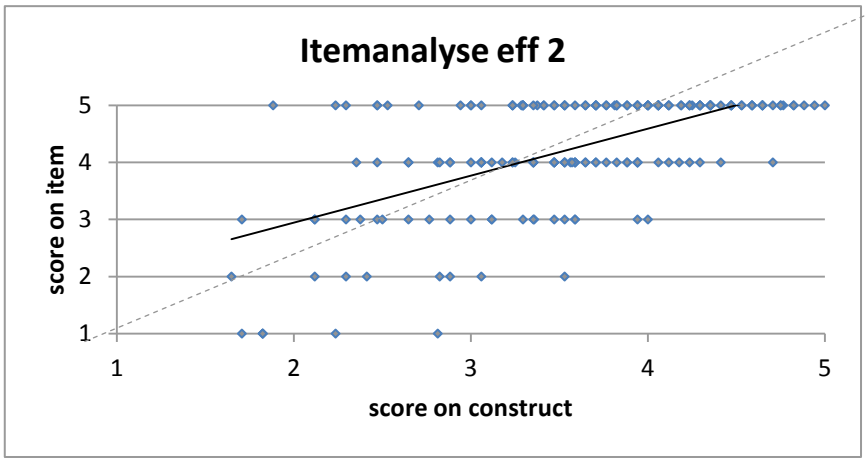


Abbildung 6.6: Die Trendlinie (schwarz, durchgezogen) liegt über der Diagonale (grau strichliert). Das Item überschätzt das Konstrukt.

Dieses Item (eff 2) überschätzt das Konstrukt allerding gleichmäßig für eine geringe und eine hohe Merkmalsausprägung, da die Linien (fast) parallel sind. Wörtlich lautet dieses Item: „Ich habe mich angestrengt diese Tätigkeit gut zu machen“. Man könnte diese Diskrepanz dahin gehend interpretieren, dass die Selbsteinschätzung der Schüler/innen meinen lässt, sich angestrengt zu haben, auch wenn sie andere Dinge, z.B. die nach der aufgewendeten Zeit, nicht mit einer hohen Ausprägung beantwortet haben.

Beispiel für ein das Konstrukt unterschätzendes Item (Abbildung 6.7):

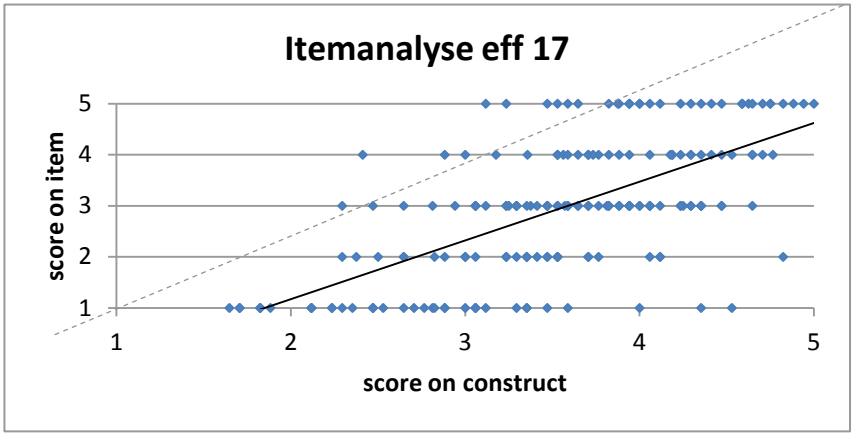


Abbildung 6.7: Die Trendlinie (schwarz, durchgezogen) liegt unter der Diagonale (grau strichliert). Das Item unterschätzt das Konstrukt.

Dieses Item (eff 17) unterschätzt das Konstrukt zwar, aber nicht gleichmäßig. Im Falle einer hohen Merkmalsausprägung misst es genauer, da die beiden Geraden näher beisammen sind. Inhaltlich lautet Item eff 17 „Am liebsten hätte ich noch nicht aufgehört, daran weiter zu arbeiten“. Offensichtlich hatten auch sonst weniger motivierte Schüler/innen hier das Bedürfnis noch ein wenig länger am CAPT zu arbeiten.

Basierend auf inhaltlichen Überlegungen, sowie auf den beiden hier vorgestellten formalen Kriterien wurde ein Set von Items ausgewählt, das geeignet erscheint, das Konstrukt *effort/importance* trennscharf und konstruktvalide abzubilden. Inhaltlich wurde darauf geachtet, dass redundante Items entfernt wurden und die verbleibenden Items das Konstrukt entsprechend der beiden *Construct Maps* (Tabelle 6.8 und Tabelle 6.9) möglichst breit abbilden sollten. Es sind dies die acht Items

eff 2, eff 4, eff 5, eff 9, eff 10, eff 11, eff 13 und eff 14.

Mit diesen Items wurde in weiterer Folge parallel zur eigentlichen CAPT-Studie eine dritte Pilotierung durchgeführt, um mit einer konfirmatorischen Faktorenanalyse die Skalenstruktur zu überprüfen, die mit dieser Vorgangsweise angestrebt wurde.

Insgesamt standen Daten von N = 249 Schüler/innen zur Verfügung. Die Schüler/innen stammten zum Großteil aus der Sekundarstufe 1, der Anteil an AHS-Schüler/innen lag bei 35 %. Der KMO-Test auf Stichprobeneignung ergab einen Wert von 0,90, was sehr gut ist.

6.5.5. Überlegungen zur Reliabilität

Führt man mit diesen Item und dieser Stichprobe eine Reliabilitätsanalyse auf Basis von Cronbach‘s innerer Konsistenanalyse durch, ergibt sich α = 0,86, was als hohe Reliabilität zu bewerten ist. Auch andere Formen die Reliabilitätsanalyse, wie z.B: Guttman‘s Split-Half-Formel[16] ergeben einen guten Wert von 0,87.

Eine konfirmatorische Faktorenanalyse dieser acht Items ergibt, dass sich nach dem Kaiser-Kriterium nur ein Faktor mit einem Eigenwert größer als 1 ergab, der 52,4 % der Varianz aufklären konnte. Da nur eine Komponente extrahiert werden konnte, erübrigt

[16] Es wurde hier Guttman’s Formel verwendet, da sie nicht mit der Annahme paralleler testteile arbeitet und daher präziser schätzt.

es sich, die Lösung zu rotieren. Das wird auch durch den Scree-Plot (Abbildung 6.8) deutlich, der die Eigenwerte des jeweiligen Faktors darstellt.

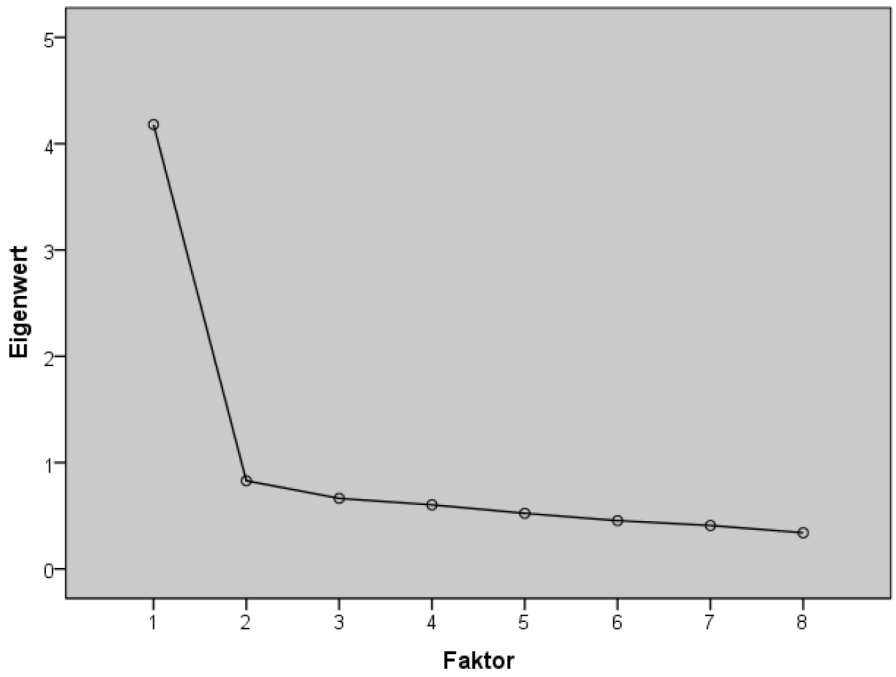


Abbildung 6.8: Scree-Plot zur Darstellung der Eigenwerte zu den acht neuen Effort-Items.

Zusammenfassend kann festgestellt werden, dass die oben angeführten acht Items eine konsistente und reliable Skala zur Messung des Konstrukts *effort/importance* (Einsatzbereitschaft/Wichtigkeit) ergeben. Diese neue Skala besteht aus drei Items des originalen IMI in Übersetzung (eff 2, eff 4, eff 5) und fünf neu entwickelten Items, die auf Basis des *Four-Building Blocks Approach* konstruiert wurden.

In weiterer Folge wurden mit den acht Items zu *effort/importance* und den Items zu den restlichen Subskalen an dem oben beschriebenen Datensatz eine konfirmatorische Faktorenanalysen durchgeführt. Das Ziel war es herauszufinden, ob die *effort/importance*-Skala, die zwar intern konsistent ist, auch bei Hinzunahme anderer Items stabil bleibt. Inhaltlich gesprochen bedeutet das herauszufinden, ob die *effort/importance*-Skala vom Konstrukt her keine Überschneidungen mit den anderen Subskalen des IMI hat. Diese würde sich in einer nicht stabilen Faktorstruktur sowie in hohen Querladungen zeigen.

Die Ergebnisse der konfirmatorischen Faktorenanalysen lassen hier für das gesamte IMI leider keine stabile Faktorstruktur erkennen. Der Großteil der acht Items zu *effort/importance* ergibt auch bei Hinzunahme anderer Items und bei gesplittetem Sample eine eigene Skala, was für die Qualität der Skala spricht. Die restlichen Items bilden zum Großteil nicht die intendierten Skalen ab. Lediglich die *relatedness*-Skala bildete in allen Analysen eine eigene, stabile Skala. Das kann dahingehend interpretiert werden, dass die r*elatedness*-Items das Konstrukt der sozialen Eingebundenheit gut repräsentieren und sich auch inhaltlich von den restlichen Items gut unterscheiden lassen. Insgesamt können somit zwei stabile Skalen identifiziert werden: die *effort/importance*-Skala mit den oben angegebenen Items und die r*elatedness*-Skala.

6.6. Abschließende Betrachtungen

Am Beginn der Arbeit wurde das gleichsam hohe wie kaum umsetzbare Ziel gesetzt, aus dem IMI ein Messinstrument zu bauen, das psychometrischen Ansprüchen genügt und das sprachlich an Schüler/innen der Sekundarstufe 1 angepasst werden kann. Die Intention dahinter war, Spielarten der Motivation, wie sie die SDT beschreibt, akkurat abbilden zu können, um sie dann mit den Ergebnissen aus den Wissenstests zu verknüpfen zu können. Darüber hinaus erschien es zu Beginn möglich, etwas über den theoretischen Rahmen des klassischen *Conceptual Change Approaches* hinauszuschauen und die Möglichkeiten von CAPT hinsichtlich eines *Conceptual Change* auch aus nicht-kognitiver Sicht zu beschreiben (Duit & Treagust, 2012).

Dieses Unterfangen kann als gescheitert angesehen werden. Das liegt zum einen daran, dass das IMI in der Operationalisierung nicht von derselben Qualität ist wie die zugehörige Motivationstheorie, die SDT (Markland & Hardy, 1997). Daher ist es nicht möglich, einfach Items und Skalen zu übernehmen und sie an den persönlichen Kontext anzupassen, wie dies die Autoren empfehlen (Deci & Ryan, 2003). Die Skalen erwiesen sich als nicht stabil in konfirmatorischen Faktorenanalysen. Zum anderen sind die Items, zumindest für Schüler/innen der Sekundarstufe 1, sprachlich nicht verständlich oder in der Nuancierung zu fein. Das hat nach einigen Versuchen des Reparierens und Umformulierens dazu geführt, dass zumindest eine Skala von Grund auf neu konstruiert wurde, nämlich die *effort/importance* Skala. Mit dieser Skala wurde begonnen, da sie am

ehesten als Prädiktor für Lernleistungen geeignet erscheint. Will man die weiteren Subskalen des IMI zur psychometrischen Beschreibung verwenden, so erscheint diese Vorgangsweise, abgesehen von der *relatedness*-Skala, ebenso angebracht zu sein und eine völlige Neukonstruktion der Items ist ernsthaft in Erwägung zu ziehen.

Die Neukonstruktion der Items nach der Methode der vier Building Blocks (Wilson, 2005) funktioniert aus praktischer Sicht gut. Aus theoretischer Sicht hat sie den Vorteil, dass Konstrukte in ihrer vollen Breite abgebildet und auf Evidenzen zurückgeführt werden können. Die Vorgangsweise beinhaltet zusätzlich, dass Aspekte der Validität bereits in der Itemkonstruktion abgehandelt werden. Sie ist aus Sicht der Autorin fruchtbar und unbedingt empfehlenswert.

Was die Beschreibung von CAPT als Unterrichtsmethode angeht, ist neben einer kognitiven auch die Beschreibung der nicht-kognitiven Einflüsse nötig, um die ganze Bandbreite an Vorgängen erfassen zu können, die sich hier abspielen. Ein zentraler Aspekt dabei ist es, entscheiden zu können, ob und in welchem Ausmaß CAPT motivierend auf die Schüler/innen wirkt. Dabei ist es von Interesse, neben der Quantität der Motivation auch ihre Qualität erfassen zu können, da ein Ziel von Unterricht und in weiterer Folge von selbstbestimmtem Lernen ist, Internalisierungsprozesse von extrinsischer zu intrinsischer Motivation zu fördern. Zu all dem ist es wichtig, ein reliables und gut funktionierendes Messinstrument zu haben, weswegen die Autorin findet, dass Entwicklungsarbeit in diese Richtung erforderlich ist, auch wenn sie bisher nicht als abgeschlossen betrachtet werden kann. Nicht abgeschlossen ist sie einerseits deshalb, weil die Entwicklung weiterer Skalen noch ausständig ist, die zur Abbildung der intrinsischen Motivation nötig sind, andererseits deshalb, weil die Itemkonstruktion und Erprobung in mehreren Zyklen erfolgen muss, bis man ein einwandfreies Messinstrument erhält. Weitere Testungen und allenfalls Verbesserungen stehen für die *effort/importance*-Skala noch aus, sind aber in vollem Umfang erst sinnvoll, wenn die Items zu den restlichen Skalen entwickelt sind. In dieser Arbeit wurde somit ein Weg vorgezeichnet, der in Folgestudien dazu seine Fortsetzung finden sollte.

7. Ergebnisse

Dieses Kapitel wird, entsprechend der Teilstudien, die zum Thema CAPT durchgeführt wurden, in Ergebnisse zur Elektrizitätslehre und Ergebnisse zur Optik geteilt. Jede der Teilstudien hatte einen eigenen Forschungsschwerpunkt und dementsprechend auch unterschiedliche Forschungsfragen (siehe Kapitel 3.2), die durch verschiedene statistische Zugänge, die der Beantwortung der Forschungsfragen im Kontext des jeweiligen Samples adäquat erschienen, bearbeitet wurden.

7.1. Ergebnisse aus der Elektrizitätslehre

Die in diesem Kapitel präsentierten Auswertungen[17] zur Entwicklung des Wissens in der Elektrizitätslehre beziehen sich allesamt auf das im Kapitel 5.1 beschriebene *Testinstrument zum Verständnis in der Elektrizitätslehre* (Urban-Woldron & Hopf, 2012). Aus diesem Testinstrument wurden die Items 1 bis 5 (teilweise zweistufig) zur Auswertung herangezogen, da diese mit den adressierten Konzepten korrespondierten. Daher ergibt sich für die Wissenstests ein möglicher Maximalscore von 9 Punkten.
Die ergänzend dazu ausgewerteten demografischen Daten wurden parallel erhoben.

7.1.1. Beschreibung des Samples

Insgesamt waren am Projekt *Cross-Age Peer Tutoring in Physik*, in dessen Rahmen die hier präsentierten Untersuchungen durchgeführt wurden, etwa 400 Kinder und Jugendliche beteiligt. Die Altersstufen der beteiligten Schüler/innen waren breit gestreut, beginnend mit Vorschulkindern (5 Jahre), über Grundschulkinder (6 bis 10 Jahre) bis hin zu Schüler/innen der Sekundarstufen 1 und 2 (10 bis 17 Jahre). Im Zentrum der Forschungsfragen (siehe Kapitel 3) und der quantitativen Datenauswertungen standen für diese Studie allerdings nur Schüler/innen aus der Sekundarstufe 1 (10 bis 14-Jährige), in ihren Rollen als Tutoren und/oder Tutees, wie in den Kapiteln 5 (Messinstrumente) und 3 (Forschungsfragen) ausführlich begründet wurde.

[17] Teile dieser Ergebnisse wurden bereits publiziert (Korner & Hopf, 2014).

Tabelle 7.1 und Abbildung 7.1 geben einen Überblick über alle jene N = 172 Schüler/innen der Sekundarstufe 1 und deren Aufteilung auf die verschiedenen Klassen, deren Daten im ersten Studienjahr ausgewertet wurden.

Klasse	Schülerzahl
Klasse 1a	17
Klasse 2	24
Klasse 3	18
Klasse 4	19
Klasse 6	22
Klasse 7	19
Klasse 8	21
Klasse 9	16
Klasse 10	16
Gesamt	**172**

Tabelle 7.1: Klassen und jeweilige Anzahl der Schüler/innen

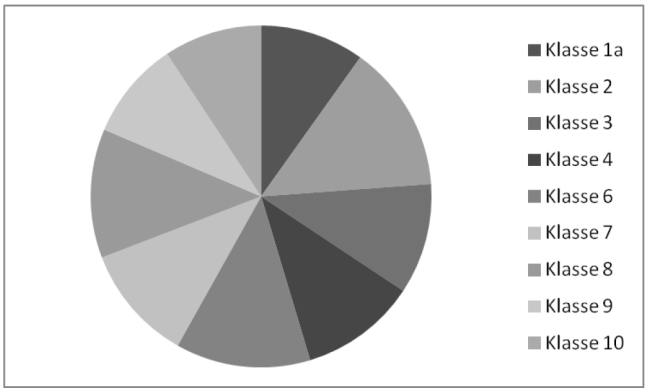


Abbildung 7.1: Klassen und Anzahl der Schüler/innen

Abbildung 7.2 zeigt den Anteil der Jungen (63 %) und der Mädchen (38 %) im Sample. Diese durchaus verzerrte Verteilung zugunsten der Jungen ist offensichtlich dem Schulsystem immanent, wie der Vergleich mit Daten aus ganz Österreich zeigt (Statistik-Austria, 2013): Das untersuchte Sample bestand aus Hauptschulklassen, mit Ausnahme von Klasse 6, die eine AHS-Klasse[18] war. Bundesweit, also in der gesamten Population, ist der Anteil der männlichen Schüler in der Hauptschule höher als der Anteil der weiblichen. Aus den zur Verfügung stehenden Daten der Statistik Austria (2013) folgernd lag im Schuljahr 2011/12 der Anteil der weiblichen Schülerinnen in der Hauptschule bei 47,5 %. Speziell im Bundesland Wien war dieser Trend noch ausgeprägter, hier lag der Mädchenanteil im selben Schuljahr bei 45,8 %. Ohne die möglichen Ursachen dafür zu erforschen, kann festgestellt werden, dass Jungen in der Hauptschule generell überrepräsentiert sind. Das ist ein Trend, der sich bis in die 1970er-Jahren zurückverfolgen lässt.

[18] AHS (Allgemein bildende höhere Schule) und Hauptschule sind unterschiedliche Schultypen in Österreich (vgl. 4.2).

Im Vergleich dazu sieht die Situation in der AHS-Unterstufe anders aus, hier lag für das Schuljahr 2011/12 bundesweit, ebenso wie wienweit, der Anteil der Mädchen bei etwa 51,8 % (Statistik-Austria, 2013).

Im Gegensatz dazu lag im hier erforschten Sample der Anteil der Mädchen in der (AHS-) Klasse 6 bei lediglich 30 %, was darauf zurückgeführt werden kann, dass es sich hier um ein naturwissenschaftliches Realgymnasium handelt. Bei diesem speziellen AHS-Typ ist der Jungenanteil, entsprechend den Interessen am naturwissenschaftlichen Unterricht (Häußler, et al., 1998), traditionell sehr hoch.

Beide Faktoren, der österreichweite Trend zu mehr Jungen in der Hauptschule und der spezielle, untersuchte Schultyp in der AHS, erklären zufriedenstellend die beobachtete Mädchen-Jungen-Verteilung im Sample (Abbildung 7.2).

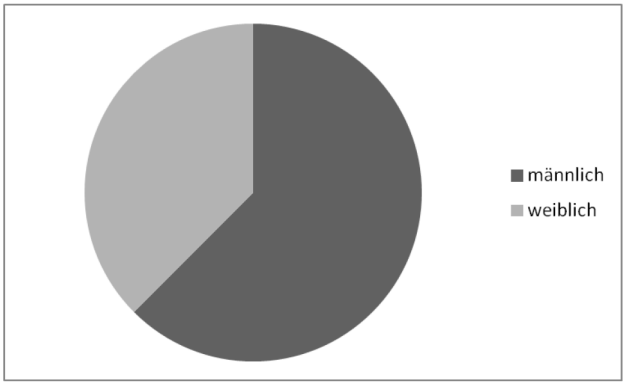


Abbildung 7.2: Anteil der männlichen und weiblichen Schüler/innen

Abbildung 7.3 zeigt die Anteile der Schüler/innen nach ihrer Muttersprache. Es wurde erhoben, ob die Muttersprache Deutsch ist (68 % aller Schüler/innen), oder ob daheim eine andere Sprache gesprochen wird (32 %). Die einzelnen Muttersprachen waren dabei nicht von Interesse, sondern ob die Sprache, in der die Instruktion stattfand, mit der Muttersprache übereinstimmt, oder nicht. Das erscheint insofern interessant, da beim Peer Tutoring die Tutoren *erklären* müssen und daher neben ihrem konzeptuellen Wissen und ihren diagnostischen Fähigkeiten stark auf ihre sprachlichen Kenntnisse angewiesen sind. Vice versa müssen die Tutees auf rein sprachlicher Ebene zuerst verstehen können, was von ihnen verlangt wird, bevor eine konzeptuelle Weiterentwicklung stattfinden kann.

Die Abbildung 7.3 kann aber keine Auskunft darüber geben, wie sich die Muttersprachen innerhalb der Klassen verteilen und ist daher mit einiger Vorsicht zu interpretieren. Im aktuellen Fall lag der Anteil der Schüler/innen mit nicht-deutscher Muttersprache in Klasse 2 bei 80 %, während er in Klasse 8 bei 0 % lag. In den folgenden Analysen stehen entsprechend den Forschungsfragen die Zusammenhänge zwischen Muttersprache und Posttestergebnis nicht im Mittelpunkt, da der Fokus auf dem Einfluss der Rollen liegt. Dennoch wurde genau untersucht, wie stark sich Schüler/innen mit deutscher Muttersprache im Vorwissen von denen mit einer anderen Muttersprache unterscheiden, um in einem abschließenden multilinearen Regressionsmodell eine mögliche Bedeutung dieses Parameters untersuchen zu können.

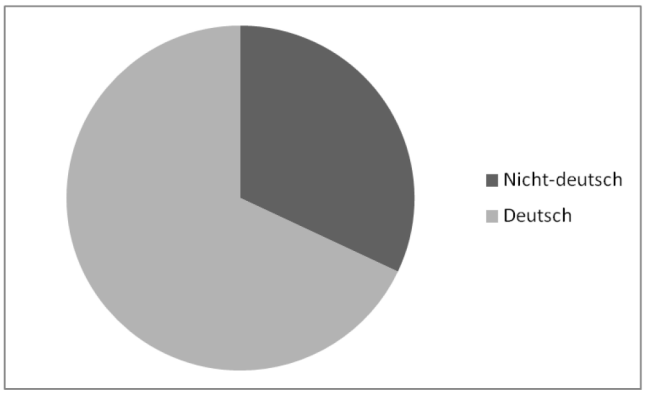


Abbildung 7.3: Anteil der Schüler/innen mit deutscher und mit nicht-deutscher Muttersprache

Da im Zentrum der Forschungsfragen 2 und 3 der mögliche Einfluss der Rollen steht, soll Abbildung 7.4 einen Überblick über die Rollenverteilung geben. *Tutoren* sind jene Schüler/innen, die im Rahmen des Tutorings aktiv ihr Wissen weitergeben, *Tutees* sind jene in der passiven Rolle, die durch die Tutoren das Tutoring empfangen. Als *Tutees und Tutoren* werden Schüler/innen bezeichnet, die zuerst ein Tutoring erhalten haben und nach einem eigenen Mentoring mit einer weiteren Klasse ein Tutoring als Tutoren durchführten (vgl. Forschungsdesign, S 48).

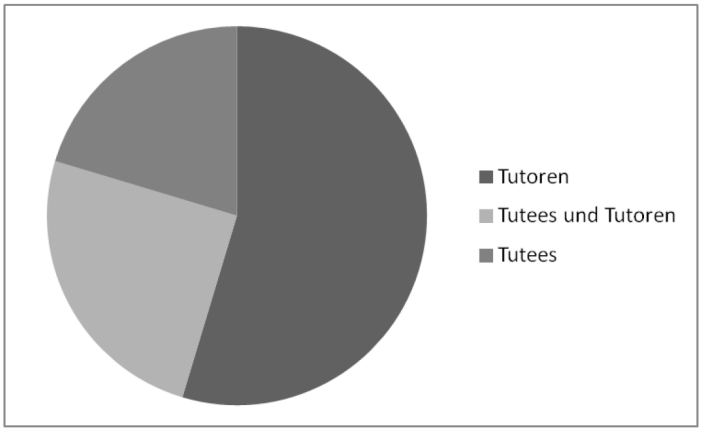


Abbildung 7.4: Verteilung der Rollen innerhalb des Tutoringprozesses

Obwohl zu erwarten gewesen wäre, dass die Anzahl der Tutor/innen genau der der Tutees entspricht, trifft das im vorgestellten Sample nicht zu. Die Tutoren bilden im Rahmen der weiteren Analysen die Mehrheit, da ein Teil der Schüler/innen, die als Tutees am Prozess beteiligt waren, nicht aus der Sekundarstufe 1 stammten, sondern aus der Volksschule und daher in der vorliegenden Studie nicht berücksichtigt wurden. Die genaue Verteilung ist Tabelle 7.2 zu entnehmen.

Rolle	Anteil
Tutoren	55 %
Tutees	25 %
Tutee/Tutoren	20 %

Tabelle 7.2: Anteile der Schüler/innen nach Rolle aufgeteilt

7.1.2. Vorwissen und Zuordnung der Rollen

In diesem Kapitel sollen Ergebnisse zu Vergleichen zwischen dem Vorwissen, das die Schüler/innen mitbrachten, sowie zu der Zuteilung zu den drei verschiedenen Rollen, die es im Tutoringprozess gab, dokumentiert werden. Abgesehen von den demografischen Parametern, die im vorangehenden Kapitel beschrieben wurden, ist für die Aussagekraft der anschließenden Prae-Post – Vergleiche von entscheidendem Interesse, wie homogen dieses Sample hinsichtlich des Vorwissens und der Rollenzuteilungen ist.

Die folgenden Fragen werden in diesem Kapitel beantwortet um herauszufinden, ob es zwischen Schüler/innen aus verschiedenen Gruppen Unterschiede gab.

1. Gibt es Unterschiede im Vorwissen zur Elektrizitätslehre zwischen den Klassen?
2. Gibt es Unterschiede im Vorwissen zur Elektrizitätslehre zwischen den Klassen der beiden Schulformen Hauptschule und AHS?
3. Gibt es Unterschiede im Vorwissen zur Elektrizitätslehre, die auf die unterschiedlichen Schulstufen der Schüler/innen zurückgeführt werden können?
4. Erfolgte die Rollenzuteilung (*Tutees*, *Tutoren*, *Tutees/Tutoren*) wirklich zufällig, wie das durch das Versuchsdesign intendiert war?

Unterschiedliche Voraussetzungen, die sich mit den Praetests abbilden ließen, wären insofern nicht unerwartet gewesen, bestand das Sample doch aus acht Hauptschulklassen und einer AHS-Klasse, die drei verschiedenen Schulstufen (6, 7 und 8) angehörten. Zudem wiesen die Klassen stark unterschiedliche Anteile an Schüler/innen mit nicht-deutscher Muttersprache auf. Diese Faktoren könnten sich eventuell auf das Vorwissen ausgewirkt haben. In diesem Fall wäre über den Ablauf des Tutoringprozesses noch einmal nachzudenken gewesen, das Forschungsdesign zu adaptieren und Analysen eher auf Klassenebene statt über das ganze Sample hinweg anzustellen gewesen.

Außerdem wollten wir ausschließen, dass, trotz gegenteiliger Intention, die leistungsstärkeren Klassen als Tutoren ausgewählt worden waren, z.B. weil diese von ihren Lehrer/innen eher vorgeschlagen worden waren. Dann wäre es nicht verwunderlich gewesen, wenn Tutoren schon allein aufgrund dieser Tatsache und nicht wegen der Intervention die besseren Posttest-Ergebnisse gezeigt hätten. Zudem haben sich im Zuge der Auswertungen zur Optik Hinweise darauf ergeben, dass der Tutoringprozess erst ab einem gewissen kognitiven Mindestniveau Wirkung zeigt. Dieser Effekt wäre im Fall ausschließlich leistungsstarker Schüler/innen in der Tutorenrolle nicht identifizierbar.

Könnte man diese beiden Einwände allerdings mit Hilfe der Daten entkräften, hätte man ein statistisch aussagekräftiges Sample, auf Basis dessen allgemein gültige Folgerungen über mögliche Unterschiede in Lernwirksamkeit zwischen den Rollen gemacht werden können. Aus diesem Grund wurden die nachfolgenden Analysen durchgeführt.

***Testhypothese zu Frage 1**: Die einzelnen Klassen unterschieden sich hinsichtlich der Praetest-Ergebnisse nicht voneinander.*

Die Hypothese wurde mithilfe einer einfaktoriellen Varianzanalyse ohne Messwiederholung getestet. Tabelle 7.3 gibt einen Überblick über die erzielten Mittelwerte des Praetests im Klassenvergleich an.

	N	Mittelwert	SD
Klasse 1a	17	4,94	3,42
Klasse 2	24	4,21	1,91
Klasse 3	18	4,94	2,75
Klasse 4	19	4,63	2,67
Klasse 6	22	4,45	3,41
Klasse 7	19	3,11	2,23
Klasse 8	21	4,19	2,48
Klasse 9	16	4,25	1,88
Klasse 10	16	5,00	2,25
Gesamt	172	4,39	2,61

Tabelle 7.3: Mittelwerte der Praetest-Ergebnisse im Vergleich

Die Voraussetzungen, die das Sample erfüllen muss, damit eine derartige Analyse durchgeführt werden kann, wurden vor Durchführung der ANOVA überprüft:

- Normalverteilung der abhängigen Variablen in der Grundgesamtheit
- Varianzhomogenität
- Unabhängigkeit der Beobachtungen
- Intervallskalenniveau der anhängigen Variablen

Die Normalverteilungsannahme kann nur im Sample getestet werden. Mit Hilfe von Abbildung 7.5 lässt sich beurteilen, wie gut diese Annahme im Sample zutrifft. Die hier dargestellte Verteilung ist zwar ausreichend symmetrisch (Schiefe = 0,043), aber doch recht breitgipfelig (Kurtosis = – 0,944). Im Sample ist die Normalverteilungsannahme nur mäßig gut erfüllt. Allerdings sollte aufgrund der wesentlich größeren Zahl an Personen diese Annahme für die Grundgesamtheit gut zutreffen. Falls dem nicht so wäre, ist

jedoch die ANOVA stabil gegenüber dieser Verletzung der Voraussetzung. Zur Absicherung des Ergebnisses wurde noch ein robuster Test (Welch-Test) durchgeführt.

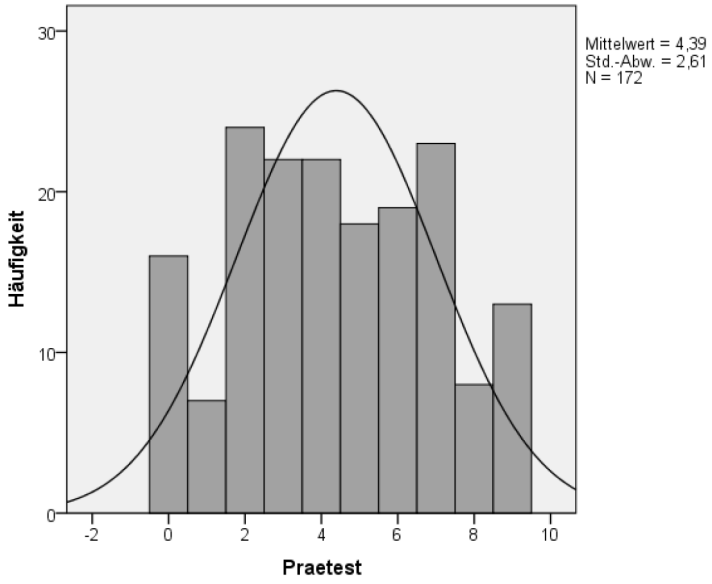


Abbildung 7.5: Histogramm der Variablen Praetest im Vergleich zur Normalverteilung

Die Normalverteilungsannahme kann nur im Sample getestet werden. Mit Hilfe von Abbildung 7.5 lässt sich beurteilen, wie gut diese Annahme im Sample zutrifft. Die hier dargestellte Verteilung ist zwar ausreichend symmetrisch (Schiefe = 0,043), aber doch recht breitgipfelig (Kurtosis = – 0,944). Im Sample ist die Normalverteilungsannahme nur mäßig gut erfüllt. Allerdings sollte aufgrund der wesentlich größeren Zahl an Personen diese Annahme für die Grundgesamtheit gut zutreffen. Falls dem nicht so wäre, ist jedoch die ANOVA stabil gegenüber dieser Verletzung der Voraussetzung. Zur Absicherung des Ergebnisses wurde noch ein robuster Test (Welch-Test) durchgeführt.

Die Annahme der Varianzhomogenität kann aufgrund des hochsignifikanten Levene-Tests ($p < 0,001$) nicht aufrechterhalten werden. Da die ANOVA auch gegenüber der Verletzung dieser Voraussetzung robust ist, genügt es hier einen F_{max}-Test durchzuführen. Dabei wird der Quotient zwischen der größten und der kleinsten Gruppenvarianz gebildet. Falls die Gruppengrößen etwa gleich groß sind (maximal 4:1), was hier der Fall ist, wird empfohlen, dass dieser Quotient kleiner als 10 sein soll (Tabachnik & Fidell, 2006). Im vorliegenden Fall ist $F_{max} = 3,42^2 / 1,88^2 = 3,3 < 10$ und ein F-Test auf einem Signifikanzniveau von $\alpha = 0,05$ kann durchgeführt und ohne

Korrekturen interpretiert werden. Andernfalls wäre die Irrtumswahrscheinlichkeit, fälschlicherweise die Nullhypothese abzulehnen, von p = 0,05 zu halbieren.

Levene-Statistik	df1	df2	Signifikanz
3,726	8	163	0,000

Tabelle 7.4: Der Levene-Test auf Homogenität der Varianzen fällt hochsignifikant aus.

Die beiden letzten Voraussetzungen sind die Unabhängigkeit der Testergebnisse und das Intervallskalenniveau der Messwerte. Letzteres ist durch die übliche Bewertung durch Punkte erfüllt. Gegenüber Verletzungen der Unabhängigkeitsvoraussetzung ist die Varianzanalyse sehr empfindlich. Mit der, bei diesen Tests üblichen, Vorgangsweise (Einzelarbeit ohne vorherige Instruktion) sollte die Unabhängigkeit jedoch gegeben sein. Diese Einschätzung unterstützen auch die Analysen der Praetests, wonach es zwischen den Klassen keine signifikanten Unterschiede gab.

Die Ergebnisse der eigentlichen Analyse, der ANOVA über die Klassen hinweg, sind in Tabelle 7.5 dargestellt. Mit einem empirischen F-Wert von F (8,163) = 0,936 auf einem Signifikanzniveau von p = 0,489 kann die Testhypothese 1 daher nicht abgelehnt werden. Es ist daher nicht zu erwarten, dass in der Population das Vorwissen bezüglich der getesteten Konzepte der Elektrizitätslehre in dieser Altersstufe unterschiedlich ist.

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	51,154	8	6,394	0,936	**0,489**
Innerhalb der Gruppen	1113,747	163	6,833		
Gesamt	1164,901	171			

Tabelle 7.5: Die ANOVA zeigt keine signifikanten Unterschiede zwischen den Schülergruppen.

Parallel zur ANOVA wurde ein robustes Testverfahren eingesetzt (Tabelle 7.6), das diese strengen Voraussetzungen nicht benötigt: Der Welch-Test bestätigt das vorherige Ergebnis, dass sich Schüler/innen dieser Altersstufe nicht signifikant im Vorwissen unterschieden.

Robuste Testverfahren zur Prüfung auf Gleichheit der Mittelwerte				
Praetest				
	Statistik[a]	df1	df2	Sig.
Welch-Test	1,052	8	66,084	0,407

a. Asymptotisch F-verteilt

Tabelle 7.6: Absicherung der Ergebnisse der ANOVA durch einen Welch-Test.

Dieses Ergebnis ist insofern interessant, da eine Klasse (1a) im Jahr zuvor bereits Unterricht in Elektrizitätslehre erhalten hatte. Aufgrund der Praetests, die sich für diese Klasse nicht signifikant von denen der restlichen Klassen unterschieden, konnte diese Klasse trotzdem im Sample behalten werden. Anders gesagt: Klasse 1a schnitt etwa ein Jahr nach dem Unterricht in Elektrizitätslehre nicht besser (auch nicht schlechter!) ab, als jene, die noch nie darin unterrichtet worden waren.

Der zweite Punkt, hinsichtlich dessen das Sample auf Homogenität getestet werden soll, ist ein möglicher Unterschied zwischen den Schüler/innen der Schulformen Hauptschule und AHS.

Testhypothese zu Frage 2: *Die Praetest-Ergebnisse einzelnen Schüler/innen unterscheiden sich hinsichtlich der Schulform nicht voneinander.*

	N	Mittelwert	SD
HS	150	4,38	2,49
AHS	22	4,45	3,41

Tabelle 7.7: Praetest-Ergebnisse nach Schulformen (Hauptschule, AHS)

Tabelle 7.7 zeigt die mittleren Punktescores der Praetests von Hauptschüler/innen im Vergleich zu denen der AHS. Ein t-Test, der die beiden Mittelwertunterschiede entsprechend der Testhypothese testet, liefert auch bei angenommener Ungleichheit der Varianzen kein signifikantes Ergebnis: t = – 0,99 mit den korrigierten Freiheitsgraden df = 24,393 und einer Übertretungswahrscheinlichkeit p = 0,922. Somit kann diese Testhypothese nicht abgelehnt werden und man kann davon ausgehen, dass sich auch in der Grundgesamtheit die Wissensvoraussetzungen zwischen AHS- und Hauptschüler/innen nicht unterschieden.

Die dritte Fragestellung betrifft die Auswahl der Schüler/innen für die einzelnen Rollen innerhalb des Tutoringprozesses. Es war beabsichtigt, dass diese Zuordnung nicht auf

Basis von Empfehlungen oder kognitiven Fähigkeiten erfolgen sollte, sondern zufällig.

Testhypothese zu Frage 3: *Die Praetest-Ergebnisse der Schüler/innen unterscheiden sich im Hinblick auf die unterschiedlichen Schulstufen nicht voneinander.*

Die Schüler/innen, die hier getestet wurden, stammten aus den Schulstufen 6, 7 und 8. Es wurde nach dem Prüfen der Voraussetzungen, was in diesem Fall homogene Varianzen lieferte, eine ANOVA durchgeführt. Tabelle 7.8 gibt einen Überblick über die Praetest-Ergebnisse. Es zeigt sich, dass mit steigender Schulstufe die Praetest-Ergebnisse etwas besser ausfallen.

Abhängige Variable: Praetest			
Schulstufe	Mittelwert	SD	N
6	4,11	2,38	82
7	4,58	2,65	73
8	4,94	3,42	17
Gesamt	4,39	2,61	172

Tabelle 7.8: Praetest-Ergebnisse der unterschiedlichen Schulstufen

Die ANOVA ist hier trotzdem nicht signifikant: $F\ (2, 169) = 1{,}036$ mit einer Übertretungswahrscheinlichkeit $p = 0{,}357$. Daher kann die Nullhypothese nicht abgelehnt werden und man kann von einem homogenen Vorwissen zwischen den Schulstufen ausgehen.

Testhypothese zu Frage 4*: Die Praetest-Ergebnisse der Schüler/innen hinsichtlich der drei Rollen Tutees, Tutoren und Tutees/Tutoren unterschieden sich nicht.*

Aus Tabelle 7.9 lässt sich zwar erkennen, dass die Tutees etwas schlechter abschneiden als der Rest der Schüler/innen, dass sie aber auch die kleinste Gruppe sind.

Abhängige Variable: Praetest			
Rolle	Mittelwert	SD	N
1 Tutoren	4,70	2,57	94
2 Tutees/Tutoren	4,33	2,96	43
3 Tutees	3,63	2,13	35
Gesamt	4,39	2,61	172

Tabelle 7.9: Praetest-Ergebnisse nach den unterschiedlichen Rollen im Tutoringprozess.

Daher zeigte nach Prüfen der Voraussetzungen für einen solchen Test die ANOVA keine signifikanten Unterschiede zwischen den Gruppen: F (2, 169) = 2,205 mit p = 0,113. Die Testhypothese kann somit nicht verworfen werden und es lässt sich folgern, dass die Zuteilung der Klassen zu den Rollen tatsächlich zufällig erfolgte.

Zusammenfassend kann davon ausgegangen werden, dass das untersuchte Sample hinsichtlich des Vorwissens, das die Schüler/innen vor der Intervention durch CAPT hatten, als homogen angesehen werden kann. Weder beeinflussten die unterschiedlichen Schulstandorte und Schulstufen, noch die verschiedenen Schulformen (AHS und HS) die Praetest-Ergebnisse signifikant. Auch die Zuordnung der Schüler/innen zu den drei Rollen *Tutees*, *Tutoren* und *Tutees/Tutoren* ist zufällig verteilt.

7.1.3. Prae- und Posttests im Vergleich

Nachdem im vorangehenden Kapitel die mitgebrachten Voraussetzungen der Schüler/innen geklärt wurden, werden in diesem Kapitel jene Forschungsfragen beantwortet, die sich auf die Prae- Post-Vergleiche im Zusammenhang mit dem Thema Elektrizitätslehre beziehen. Es handelt sich hier um die Forschungsfragen 1 bis 4. Zur besseren Orientierung werden diese Fragen jeweils am Beginn der Analysen wiederholt.

***Forschungsfrage 1**: Weisen Schüler/innen der Sekundarstufe 1 nach der CAPT-Intervention bessere kognitive Testergebnisse auf und wenn, wie stark ist der gemessene Effekt?*

Um diese Frage zu beantworten, wurden die Prae- und Posttests aller Schüler/innen verglichen. Da man davon ausgehen kann, dass die CAPT-Intervention keine Verminderung des Wissens der Schüler/innen bewirkt, sondern dieses allenfalls verbessert, konnte einseitig getestet werden. Tabelle 7.10 zeigt die deskriptive Statistik der Prae- und Posttests. Da hier mit listenweisem Fallausschluss gearbeitet wurde, reduziert sich die Anzahl der analysierten Tests, über die in diesem Kapitel berichtet wird, auf N = 164, entsprechend den Schüler/innen, die zu beiden Testzeitpunkten anwesend waren. Das bedeutet einen geringen Datenverlust, bei dem diese Methode des Fallausschlusses gerechtfertigt ist (Rost, 2007).

Statistik bei gepaarten Stichproben			
	Mittelwert	N	SD
Praetest	4,41	164	2,59
Posttest	5,84	164	2,85

Tabelle 7.10: Prae- und Posttests aller Schüler/innen im Vergleich

Der t-Test für gepaarte Stichproben (Tabelle 7.11) zeigt einen Unterschied zwischen Prae- und Posttest mit $t\ (163) = -5{,}826$ und $p < 0{,}001$. Dieser Unterschied wäre auch schon bei einer zweiseitigen Testung höchstsignifikant, und deshalb erst recht bei der einseitigen Testung.

Statistik bei gepaarten Stichproben					
	Gepaarte Differenzen		T	df	Sig. (2-seitig)
	Mittelwert	SD			
Praetest – Posttest	-1,42	3,12	-5,826	163	0,000

Tabelle 7.11: Ergebnis des t-Tests für alle Schüler/innen

Um die praktische Bedeutsamkeit dieses Effekts zu untersuchen, wurde die Effektstärke nach Cohen (1988) berechnet: $d_z = 0{,}46$, was einer Verschiebung der Mittelwerte zu den beiden Testzeitpunkten um 0,46 Standardabweichungen entspricht. Die Effektstärke gibt an, wie stark der beobachtete Effekt ist, in diesem Fall ist sie ein Maß dafür, wie viel besser die Schüler/innen im Posttest abschneiden als im Praetest.

Verwendet man die Kategorisierung nach Cohen (1988), so sind erst Effekte ab 0,5 als mittelgroß, solche ab 0,8 als groß einzustufen. Für die Fachdidaktik beschreiben Häußler et al. (1998) Effekte ab 0,67 als interessant, solche ab 0,46 immerhin als gut. Hattie (2009, S. 16) hingegen spricht im Kontext von Schule und Lernen bereits ab einer Effektstärke von 0,4 von der „*zone of desired effects*“, also von mittelgroßen Effekten, und klassifiziert Effekte von mehr als 0,6 als groß.

Die Teststärke, die angibt, wie wahrscheinlich die Alternativhypothese zutrifft[19], dass es durch die CAPT-Intervention zu einer Verbesserung in der Testleistung kommt, liegt hier bei 0,99.

***Forschungsfrage 2**: Weisen auch die Tutoren nach der CAPT-Intervention bessere kognitive Testergebnisse auf?*

Zur Beantwortung dieser Frage wurden die Tests jener Schüler/innen analysiert, die ausschließlich in der Tutorenrolle waren. Die Anzahl der verwertbaren Tests lag hier bei N = 91. Es konnten wieder nur die Tests jener Schüler/innen analysiert werden, die zu beiden Testzeitpunkten anwesend waren. Wiederum ging man von einem Wissenszuwachs aus, daher wird auch in diesem Falle einseitig getestet. Tabelle 7.12 gibt die deskriptiven Daten zu den Prae- und Posttests für die Gruppe der Tutoren wieder.

Statistik bei gepaarten Stichproben			
	Mittelwert	N	SD
Praetest	4,74	91	2,58
Posttest	5,98	91	2,72

Tabelle 7.12: Vergleich der Praetests und Posttests für Tutoren

Die Daten aus Tabelle 7.13 erlauben eine Beurteilung der Mittelwertunterschiede. Mit t (90) = 3,976 und p < 0,001 fällt der Unterschied in den Mittelwerten höchstsignifikant aus, insbesondere, wenn man in Betracht zieht, dass die Hypothese gerichtet war.

Statistik bei gepaarten Stichproben					
	Gepaarte Differenzen		T	df	Sig. (2-seitig)
	Mittelwert	SD			
Praetest - Posttest	-1,24	2,98	-3,976	90	0,000

Tabelle 7.13: Ergebnis des t-Tests für Tutoren

Die dabei ist die erzielte Effektstärke nach Cohen d_z = 0,42, bei einer Teststärke von gerundet[20] 1.

[19] Die Teststärke gibt an, mit welcher Wahrscheinlichkeit die Alternativhypothese zutrifft, falls sie angenommen wird. Es ist nämlich nicht davon auszugehen, dass die Alternativhypothese bloß deswegen zutrifft, weil die Nullhypothese unwahrscheinlich ist.

[20] Das verwendete Programm, G*Power 3.1, gibt 7 Nachkommastellen an, in diesem Fall 9er.

***Forschungsfrage 3**: Welche Unterschiede in den Posttest-Ergebnissen ergeben sich dadurch, dass die Schüler/innen innerhalb des Tutoring-Prozesses unterschiedliche Rollen inne hatten?*

Wie schon im Kapitel 7.1.1 ausführlich beschrieben, ist das Sample hinsichtlich der Praetests homogen. Es konnten keine signifikanten Unterschiede betreffend der Klassen, der Schulstufen oder der Schultypen festgestellt werden. Die Zuordnung der Klassen zu ihren Rollen innerhalb des Tutoringprozesses war ebenso zufällig verteilt. In weiterer Folge wurden nun die Posttests analysiert. Zeigen sich hier Unterschiede, so können diese eindeutig als Folge des Treatments betrachtet werden.

Zur Beantwortung dieser Forschungsfrage werden drei Hypothesen formuliert und getestet:

Hypothese 3.1: Es gibt Unterschiede in den Posttests, die auf die unterschiedlichen Rollen im Tutoringprozess zurückzuführen sind.

Nach einer sorgfältigen Prüfung der Voraussetzungen, die inklusive der Homogenität der Varianzen erfüllt sind, wurde eine ANOVA durchgeführt.

Tabelle 7.14 zeigt die Mittelwerte und die Standardabweichungen der drei unterschiedlichen Gruppen *Tutoren*, *Tutees/Tutoren* und *Tutees*.

Posttest			
	N	Mittelwert	SD
Tutor	92	5,99	2,70
Tutee/Tutor	38	6,76	2,51
Tutee	35	4,46	3,09
Total	165	5,84	2,84

Tabelle 7.14: Vergleiche der Posttests von Schüler/innen mit unterschiedlichen Rollen im Tutoringprozess

Die Ergebnisse der ANOVA können Tabelle 7.15 entnommen werden. Es zeigt sich, dass sich die Varianzen zwischen den Gruppen von denen innerhalb der Gruppen hochsignifikant unterscheiden: $F(2, 162) = 6{,}716$ mit $p = 0{,}002$. Die Poweranalyse ergibt eine Effektstärke von 0,66 und eine Teststärke von etwa 1. Das bedeutet, dass die hier

getestete Nullhypothese des Tests, dass sich die Gruppen nicht unterscheiden, sehr unwahrscheinlich ist und die Alternativhypothese, also Hypothese 3.1, wahrscheinlich ist und angenommen werden kann.

ONEWAY ANOVA					
Posttest					
	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	101,360	2	50,680	6,716	0,002
Innerhalb der Gruppen	1222,543	162	7,547		
Gesamt	1323,903	164			

Tabelle 7.15: Ergebnisse der ANOVA

Da die Hypothese 3.1 somit nicht verworfen werden kann, ist daher davon auszugehen, dass sich nach einer CAPT-Intervention auch in der Population Unterschiede zwischen den Posttests zeigen werden, die davon abhängen, welche Rolle die Schüler/innen inne hatten.

Hypothese 3.2: Schüler/innen, die im Tutoringprozess eine aktive Rolle innehaben (*Tutoren* und *Tutees/Tutoren*), zeigen einen größeren Wissenszuwachs als die Schüler/innen in der passiven Rolle (*Tutees*).

Aufgrund der Forschungsergebnisse vorangehender Studien, kann man davon ausgehen, dass auch die Tutoren einen Wissenszuwachs zeigen (P. A. Cohen, et al., 1982; Fogarty & Wang, 1982). Darüber hinaus ist zu erwarten, dass dieser den Wissenszuwachs der Tutees sogar übertrifft (Topping, 1996). Es ist daher sinnvoll, hier eine gerichtete Hypothese zu formulieren.

Die Analyse der Daten erfolgte hier nicht durch Post-hoc-Tests, sondern theoriegeleitet mittels geplanter Kontraste, da aus der Literatur (siehe oben) eindeutige Hinweise auf einen größeren Erfolg der Schüler/innen in der aktiven Rolle existieren. Während die ANOVA zu Hypothese 3.1 zeigt, *dass* es Mittelwertunterschiede zwischen den Rollen in den Posttests gibt, sollen die geplanten Kontraste aufdecken zwischen welchen Gruppen

sie liegen[21]. Tabelle 7.16 stellt jenen Kontrast dar, der die Schüler/innen in der aktiven Rolle mit jenen der passiven vergleicht.

Kontrast Koeffizienten			
	Rolle		
Kontrast 1	Tutoren	Tutees/Tutoren	Tutees
	1	1	-2

Tabelle 7.16: Kontrast 1, der die aktive Rolle (Tutoren, Tutees/Tutoren) mit der passiven Rolle vergleicht.

Tabelle 7.17 zeigt die Analyse von Kontrast 1: t (49,83) = 3,320 bei einer Effektstärke d_z= 0,60 und einer Teststärke von 0,91. Dieser Kontrast ist hochsignifikant (p = 0,002), insbesondere im Falle einer einseitigen Testung.

Kontrasttest für Kontrast 1						
		Kontrastwert	Std.-Fehler	T	df	Signifikanz (2-seitig)
Posttest	Varianzen sind nicht gleich	3,84	1,156	3,320	49,830	0,002

Tabelle 7.17: Ergebnis des t-Tests über den Kontrast aus Tabelle 7.16

Hypothese 3.2 kann somit nicht abgelehnt werden. Es ist zu erwarten, dass auch in der Population die Schüler/innen in der aktiven Rolle im Posttest besser abschneiden als jene in der ausschließlich passiven Rolle.

Hypothese 3.3: Es gibt keine Unterschiede im Wissenszuwachs zwischen *Tutoren* und *Tutees/Tutoren*.

Der Hauptunterschied zwischen diesen beiden Gruppen liegt in der Interventionszeit. *Tutees/Tutoren* konnten mit ihren Tutees ein zusätzliches Tutoring abhalten. Ebenso wie bei Hypothese 3.2 ließ bereits ein Studium der Literatur (Robinson, et al., 2005) auf einen möglichen Ausgang des Hypothesentests schließen, nämlich dass die Dauer der Intervention keinen großen Einfluss auf den Wissenszuwachs hat. Übereinstimmend mit

[21] Die ANOVA teilt die totale Varianz in systematischen (daher aus dem Treatment folgenden) und unsystematischen Anteil auf und vergleicht diese. Mit geplanten Kontrasttests analysiert man den systematischen Anteil weitergehend bezüglich des Einflusses verschiedener Faktoren des Treatments.

diesem Befund berichtet auch Hattie (2009, S. 184) von einem eher mäßigen Einfluss der *time on task* (d = 0,38). Viel entscheidender sei die Qualität der Instruktion.

Aus diesen Gründen werden keine Dosiseffekte vermutet. Dennoch ist diese Behauptung sorgfältig zu prüfen, da die Schüler/innen in der Doppelrolle viel länger mit dem Thema Elektrizitätslehre beschäftigt waren: Sie hatten ein Tutoring, danach ihr eigenes Mentoring und anschließend führten sie ein Tutoring mit einer weiteren Klasse durch.

Tabelle 7.18 zeigt, dass mit Kontrast 2 *Tutoren* mit *Tutees/Tutoren* verglichen wurden, während die *Tutees* unberücksichtigt bleiben.

Kontrast-Koeffizienten			
Kontrast 2	Rolle		
	Tutoren	Tutees/Tutoren	Tutees
	1	-1	0

Tabelle 7.18: Kontrast 2, vergleicht die beiden aktiven Rollen

Ergebnisse eines t-Tests über diesen Kontrast wird in Tabelle 7.19 gezeigt:

t (74,072) = -1,564 mit p = 0,122. Er ist somit **nicht** signifikant und die Hypothese 3.3 kann nicht verworfen werden. Auch die berechnete Effektstärke d_z = 0,30 und die zugehörige geringe Teststärke von 0,43 untermauern diese Interpretation.

Kontrasttest für Kontrast 2						
		Kontrastwert	Std.-Fehler	T	df	Signifikanz (2-seitig)
Posttest	Varianzen sind nicht gleich	-0,77	,495	-1,564	74,072	0,122

Tabelle 7.19: Ergebnis des t-Tests über Kontrast 2 aus Tabelle 7.18

Es daher davon auszugehen, dass die Dauer der Intervention auch in der Population nicht entscheidend ist. Im Speziellen heißt das, dass die *Tutees/Tutoren* aufgrund der viel längeren Dauer der Intervention keinen signifikanten Vorsprung gegenüber den *Tutoren* haben.

Zusammenfassend kann man aus den Analysen zu allen drei Hypothesen, 3.1 bis 3.3, feststellen, dass signifikante Unterschiede in den Posttests nur auf Basis einer Einteilung des Samples nach Rollen gefunden werden können. Die Schüler/innen in der aktiven Rolle (*Tutoren*, *Tutees/Tutoren*) schneiden dabei besser ab als jene in der passiven.

Dosiseffekte aufgrund der unterschiedlichen Interventionszeiten konnten dabei keine nachgewiesen werden.

***Forschungsfrage 4**: Welche relevanten Prädiktoren können für die Modellierung der Posttest-Ergebnisse identifiziert werden?*

Die Analysen zu den Forschungsfragen 1 bis 3 zeigten vorhandene und nicht vorhandene Abhängigkeiten der Posttest-Ergebnisse von unterschiedlichen Parametern. So konnte gezeigt werden, dass durch eine CAPT Intervention nicht nur ein praktisch bedeutsamer Effekt für alle Schüler/innen zu erwarten ist, sondern dass insbesondere auch die Tutoren profitierten. In den Analysen zu den Posttests stellte sich heraus, dass lediglich der Faktor *aktive Rolle* das Sample in Untergruppen teilte, die sich hochsignifikant unterschieden. Demgegenüber ist die Interventionsdauer nicht von ausgeprägter Bedeutung.

Jene Prädiktoren, die sich im Zuge der obigen Analysen als signifikant herausstellten, nämlich das Abschneiden im Praetest und die zugeteilte Rolle, wurden in ein multiples lineares Regressionsmodell (MLR) aufgenommen. Die MLR bietet sich als Methode der Wahl an, da sich damit nicht nur die Zusammenhänge zwischen diesen Prädiktoren darstellen lassen, sondern sich auch ihre individuellen Beiträge zum Posttestergebnis quantitativ abschätzen lassen. Zusätzlich wurde mittels MLR der Einfluss der demografischen Parameter analysiert, die in dieser Untersuchung erhoben wurden und die bislang in der Datenanalyse noch keinen Niederschlag gefunden hatten. Mithilfe aller erhobenen Variablen sollte ein MLR-Modell entwickelt werden, das bei einem höchstmöglichen Maß an Signifikanz der Prädiktoren einen möglichst großen Anteil der Varianz erklären kann und dennoch möglichst einfach ist.

Um diese weiteren, möglichen Einflussfaktoren auf die Posttest-Ergebnisse identifizieren zu können, wurde innerhalb der theoretisch fundierten Parameter, die erhoben wurden, ein exploratives Vorgehen gewählt. Es wurden für diese Parameter zuerst die Korrelationen nach Pearson mit den Posttest-Ergebnissen und deren Signifikanzen ermittelt und begutachtet. Tabelle 7.20 gibt einen Überblick über die berechneten Werte. Die Variable *Ist Tutor* bezeichnet hier alle Schüler/innen, die im Rahmen des

Tutoringprozesses in der aktiven Rolle waren, daher zumindest einmal die Tutorenrolle innehatten. Sie subsumiert sowohl die *Tutoren*, als auch die *Tutees/Tutoren*. Die Variable *Muttersprache* ist so kodiert, dass der Wert 1 Deutsch als Muttersprache bezeichnet und 0 eine andere Muttersprache. Bei der Variablen *Geschlecht* wurde die Kodierung so gewählt, dass der Wert 1 weiblich bedeutet und 0 männlich. Hingegen ist die Variable *Letzte Note in Physik* entsprechend der üblichen Notenskala in Österreich von 1 (Sehr gut) bis 5 (Nicht genügend) kodiert, was die negative Korrelation erklärt. Diese Korrelation ist zwar schwach, aber signifikant.

Wie aus den Ergebnissen der Forschungsfragen 2 und 3 zu erwarten war, sind die Korrelationen des Posttests mit dem Praetest und der aktiven Rolle (*Ist Tutor*) am stärksten und auch höchstsignifikant sowie inhaltlich gut interpretierbar.

Korrelationen nach Pearson zu den Posttests		
Variable	r	Signifikanz (einseitig)
Praetest	0,426	< 0,001
Ist Tutor	0,309	< 0,001
Muttersprache	0,182	0,015
Geschlecht	-0,069	0,207
Letzte Note in Physik	-0,151	0,036

Tabelle 7.20: Korrelationen nach Pearson der angegeben Variablen zum Posttestergebnis

Die Variable *Muttersprache* zeigt zwar eine geringe Korrelation (0,182), die aber signifikant ist. Die Variable *Geschlecht* zeigt eine sehr geringe, nicht-signifikante Korrelation zu den Posttest-Ergebnissen. Das negative Vorzeichen kann so interpretiert werden, dass Mädchen einen etwas schlechteren Posttest als Jungen aufweisen.

Entsprechend dieses Befundes wurde eine hierarchische Regressionsanalyse durchgeführt. Zuvor aber wurden sorgfältig die Voraussetzungen geprüft, unter denen eine Regressionsanalyse durchgeführt und stichhaltig interpretiert werden kann.

Prüfen der Modellannahmen für eine multiple lineare Regressionsanalyse

Die Regressionsanalysen schätzen nach der Methode der *Ordinary Least Squares* (OLS)

die Regressionskoeffizienten so, dass die erhaltene Regressionsgerade die Daten bestmöglich beschreibt. Dazu müssen die folgenden Voraussetzungen erfüllt sein, damit die geschätzten Regressionskoeffizienten die *Best Linear Unbiased Estimators* (beste unverzerrte Schätzer) sind.

- Lineare Zusammenhänge zwischen den Variablen
- Homoskedastizität der Prädiktorvariablen
- keine Multikollinearität
- Keine Autokorrelation der Residuen
- Normalverteilung der Residuen

Die Testung dieser Voraussetzungen erfolgte auf Basis der Inspektion geeigneter Streudiagramme oder Histogramme. Diese Beurteilung reicht in der Regel aus, wenn sie nicht sogar Vorteile gegenüber diverser statistischer Tests hat (Bühner & Ziegler, 2009). So hängt z.B. das Ergebnis des Kolmogoroff-Smirnoff-Tests auf Normalverteilung stark von der Stichprobengröße ab oder der Durbin-Watson-Test auf Autokorrelation der Residuen von der Reihenfolge der Datenzeilen.

Die Voraussetzung der Linearität zwischen Prädiktor- und Kriteriumsvariablen lässt sich am besten durch ein Streudiagramm zeigen, wie es Abbildung 7.6 für den Prädiktor Praetest und das Kriterium Posttest darstellt.

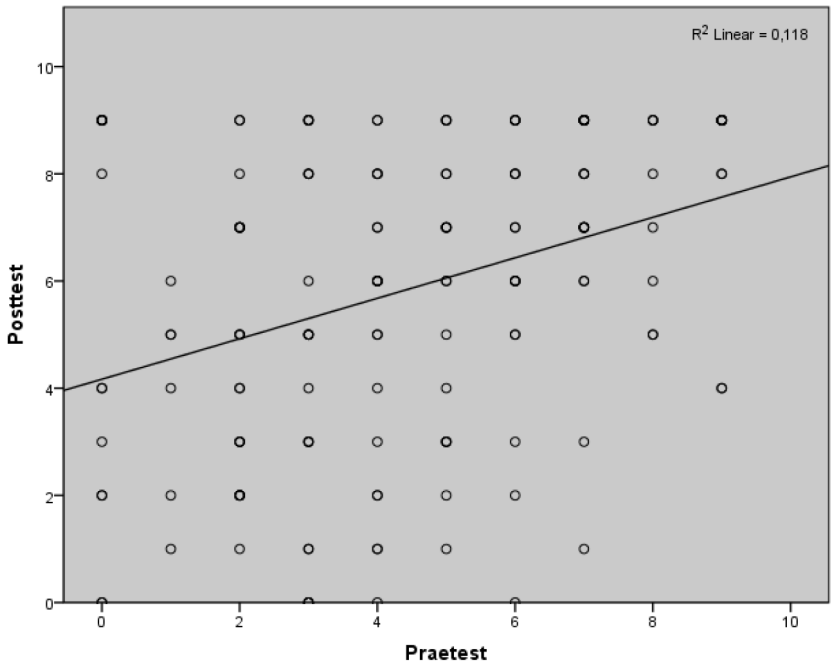


Abbildung 7.6: Überprüfung des linearen Zusammenhanges zwischen Praetest und Posttest mittels Streudiagramm als Voraussetzung für die Durchführung einer MLR

Für die weiteren Prädiktorvariablen erübrigen sich diese Diagramme, da sie dichotom sind und daher immer ein linearer Zusammenhang besteht. Lediglich für die Variable *Letzte Note in Physik* ist so ein Diagramm sinnvoll. Diese Variable stellte sich aber im Rahmen der folgenden Analysen als nicht aussagekräftig heraus, was auch schon dieses Streudiagramm (Abbildung 7.7) vermuten lässt: Die Korrelation ist zwar wie erwartet negativ („5“ ist die schlechteste Note), jedoch ist der multiple Determinationskoeffizient R^2 sehr klein.

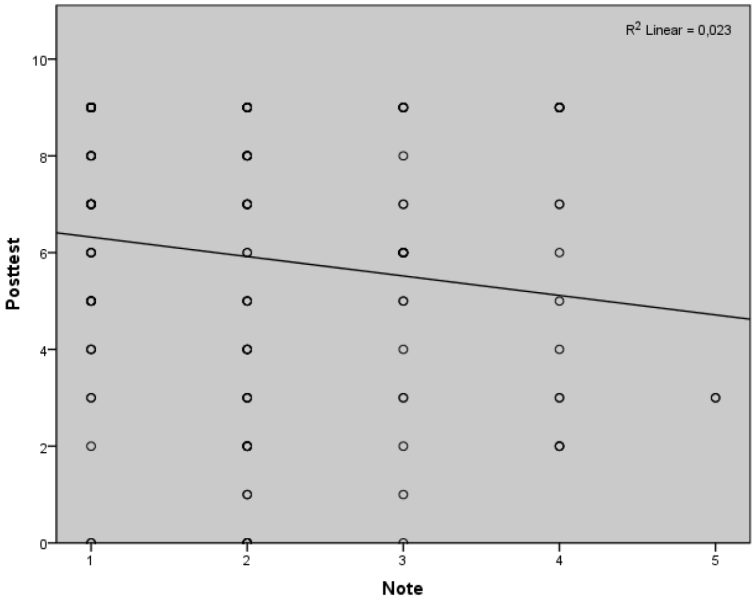


Abbildung 7.7: Streudiagramm für die Prädiktorvariable Note und die Kriteriumsvariable Posttest

Abbildung 7.8 lässt eine Bewertung der Homoskedastizität des Samples zu. Es ist auf Basis dieses Diagramms davon auszugehen, dass dieses Sample diese Voraussetzung mäßig gut erfüllt, wenn auch nicht perfekt. Die standardisierten Residuen geben die Abweichungen der beobachteten Werte zu denen wieder, die das Modell vorhersagt. Insgesamt sollten die standardisierten Residuen im Streudiagramm über alle geschätzten Werte gleich bleiben. Tatsächlich fallen im Intervall [-1;2] die vorhergesagten Werte

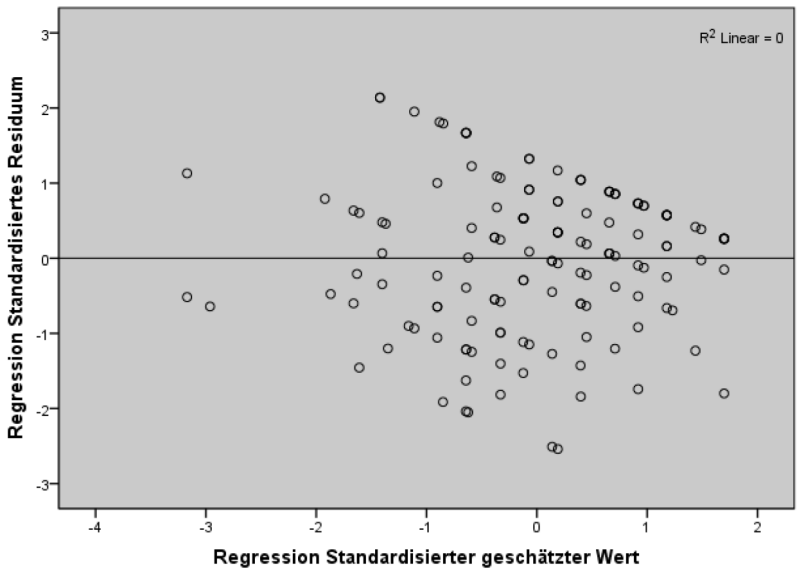


Abbildung 7.8: Streudiagramm, abhängige Variable: Posttest

etwas zu groß aus (negative Residuen). Daher kann es passieren, dass die Konfidenzintervalle der Regressionskoeffizienten ungenau geschätzt werden und man den Überschreitungswahrscheinlichkeiten (p-Werte) nicht ganz trauen kann.
Die Überprüfung, ob Multikollinearitäten vorliegen, erfolgte post hoc im Rahmen der Kollinearitätsdiagnose, deren Ergebnisse im Detail weiter unten angeführt sind. Es sei aber vorweg genommen, dass keine Kollinearitäten der Prädiktorvariablen vorlagen.

Die Überprüfung der Autokorrelationen der Residuen soll zeigen, ob die Vorhersagefehler unabhängig oder korreliert sind. Bei Autokorrelationen handelt es sich um regelmäßige Komponenten in einer Messung, was bedeutet, dass die erhobenen Werte der Variablen nicht rein zufällig sind. Das kann zum Beispiel dann passieren, wenn eine weitere Variable auf die Prädiktorvariablen wirkt, die in die Regressionsanalyse nicht aufgenommen wurde, aber einen Trend verursacht (Moderatorvariable). Wenn man die Autokorrelationen untersucht, bekommt man Aufschluss über die Unabhängigkeit der Messungen und über die Qualität des Modells an sich.
Um die Autokorrelationen zu beurteilen wurde zunächst ein Durbin-Watson-Test durchgeführt. Die Durbin-Watson-Statistik wirft prinzipiell einen Wertebereich zwischen 0 und 4 aus. Für das weiter unten vorgestellte Modell 3 lieferte sie den Wert 1,859 und für Modell 4 den Wert 1,814. Bei diesem Test ist ein Wert um 2 erstrebenswert, Werte

zwischen 1,5 und 2,5 gelten als akzeptabel (Lübbert, 1999). Somit deutet dieser Test nicht auf das Vorliegen von Autokorrelationen hin.

Zur Absicherung der Ergebnisse wurde zusätzlich für die standardisierten Residuen ein Autokorrelogramm (Abbildung 7.9) und ein partielles Autokorrelogramm (Abbildung 7.10) erstellt. Beide Diagramme stellen die Korrelationskoeffizienten der einzelnen Reihenwerte dar, die *k* Intervalle voneinander entfernt sind (Lag-Nummer *k*). Die beiden Linien stellen die obere und die untere Konfidenzgrenze des 95%-Konfidenzintervalles dar, unter der Nullhypothese, nach der keine Autokorrelationen vorliegen.

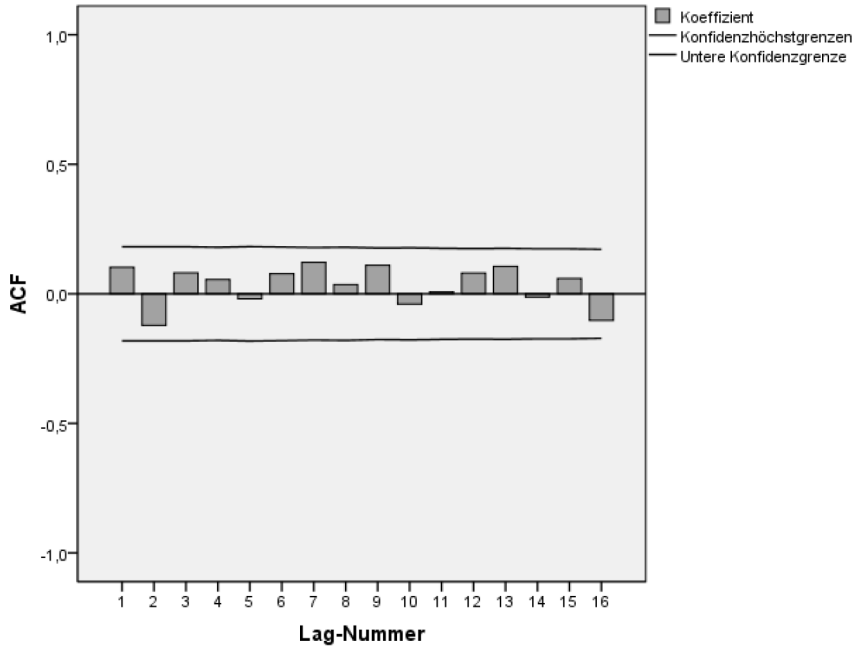


Abbildung 7.9: Autokorrelogramm für die standardisierten Residuen

Während Abbildung 7.9 die Korrelationen erster Ordnung darstellt, stellt Abbildung 7.10 die partiellen Autokorrelationen dar. Hier werden die Korrelationen größerer zeitlicher Verschiebung um jene kleinerer zeitlicher Verschiebung bereinigt (Brosius, 1998).

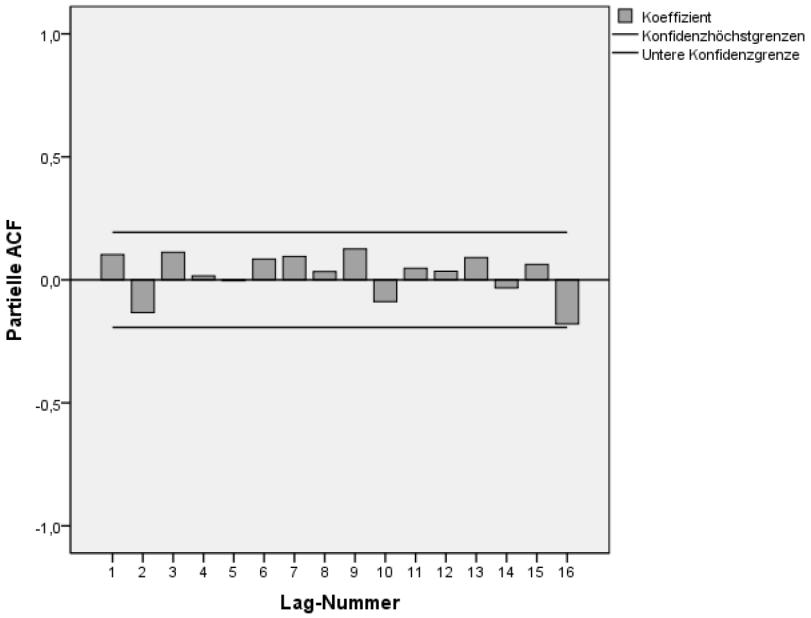


Abbildung 7.10: Partielles Autokorrelogramm für die standardisierten Residuen

Beide Diagramme lassen darauf schließen, dass keine systematischen Korrelationen der Residuen vorliegen. Die Werte sind demzufolge rein zufällig zustande gekommen. Insbesondere ist nicht zu vermuten, dass es weitere Variable gibt, die in den Modellen 3 und 4 nicht berücksichtigt wurden und die moderierend auf die dort berücksichtigten Variablen wirken.

Die letzte zu prüfende Voraussetzung, die das Sample erfüllen muss, damit eine MLR durchgeführt und aussagekräftig interpretiert werden kann, betrifft die Normalverteilung der Residuen. Abbildung 7.11 liefert einen starken Hinweis darauf, dass diese Voraussetzung erfüllt ist. Zusätzlich zeigt Abbildung 7.12 ein Probatility-Probability-Diagramm (P-P-Plot) zwischen den beobachteten und erwarteten kumulierten Wahrscheinlichkeiten. Unter der Annahme der Normalverteilung sollten die dargestellten Werte entlang einer Geraden liegen. Das bedeutet insbesondere, dass keine *fat tails* oder *thin tails* auftreten sollen, daher keine s-förmigen Abweichungen nach oben oder unten im Bereich der kleinen oder großen Wahrscheinlichkeiten. Die hier dargestellten Daten liefern eine sehr gute Übereinstimmung mit der geforderten Geraden, was wiederum auf eine gut erfüllte Normalverteilungsannahme schließen lässt.

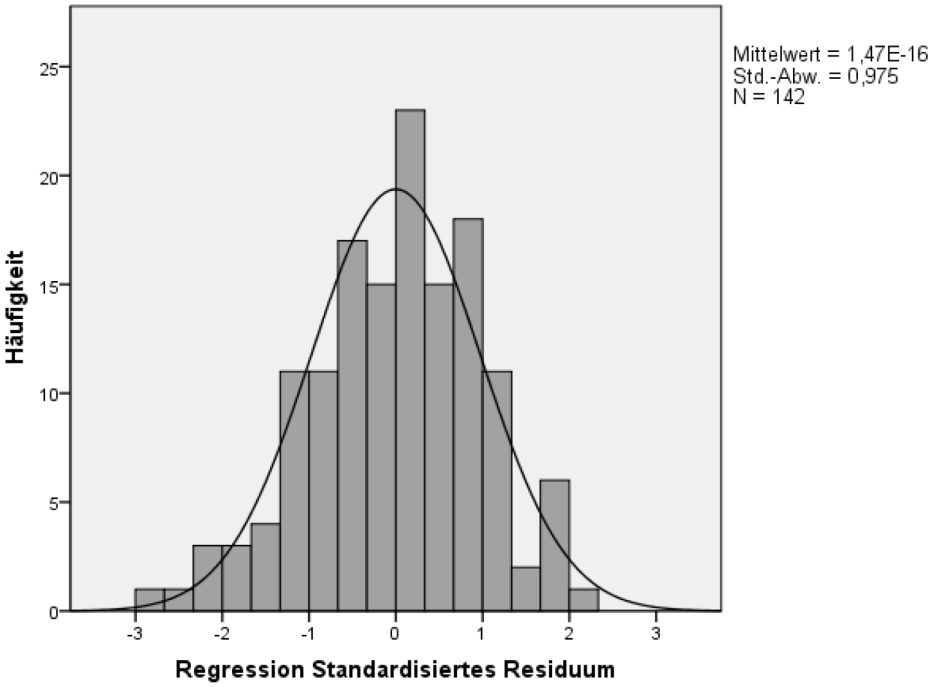


Abbildung 7.11: Histogramm, abhängige Variable: Posttest

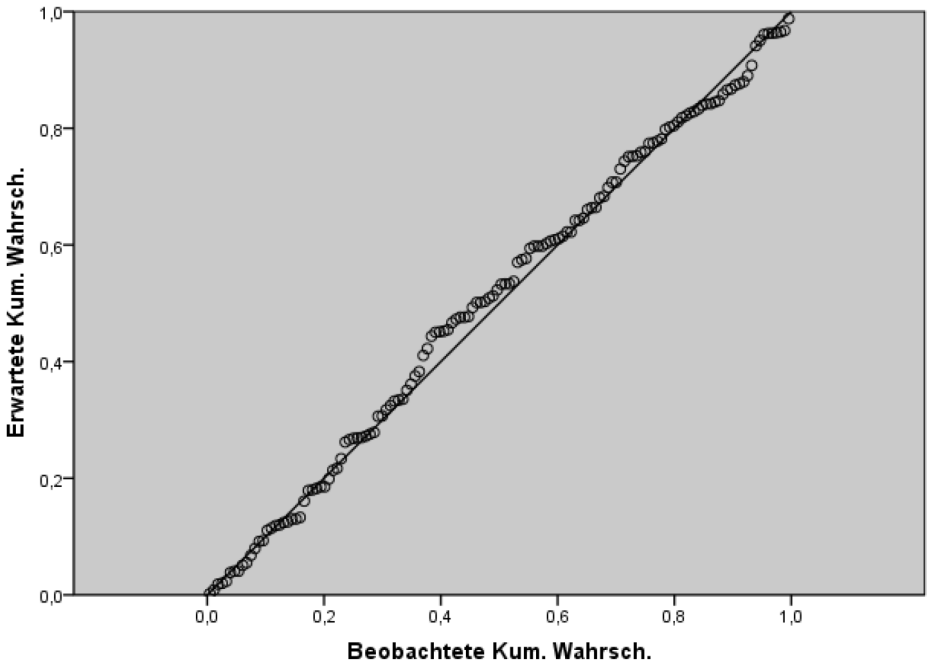


Abbildung 7.12: Probatility-Probatility-Diagramm des standardisierten Residuums für die Abhängige Variable Posttest

Die Multiplen Linearen Regressionsmodelle

Im Rahmen der Analysen wurden verschiedene Modelle getestet. Dabei stellten sich vier Modelle als berichtenswert und gut interpretierbar heraus.

Es handelt sich von der Herangehensweise her um eine theoriegeleitete hierarchische Regressionsanalyse. Für jedes der berechneten Modelle wurden zur Vorhersage der abhängigen Variable *Posttest* die Prädiktoren in Blöcken hinzugenommen. Innerhalb dieser Blöcke wurden die Prädiktoren gleichzeitig in das Modell aufgenommen. Diese Vorgangsweise läuft auch unter der Bezeichnung Einschlussmethode (forced entry) und ist sinnvoll, da es sich hier um eine Auswahl von Prädiktoren auf Basis vorangegangener Forschung handelt. Ihre Hinzunahme ist somit theoretisch fundiert und daher ist ein rein exploratives Vorgehen nicht notwendig.

Aus diesem Grund wurde auch auf die schrittweisen Methoden „vorwärts" und „rückwärts" im Rahmen dieser Analysen verzichtet[22]. Diese Methoden schließen auf Basis rein mathematischer Operationen Prädiktoren aus oder nehmen sie hinzu, ohne ihre inhaltliche Bedeutung zu berücksichtigen. Die Ergebnisse sind daher stark von den Stichprobeneigenschaften abhängig. Darüber hinaus wäre eine, im Zusammenhang mit schrittweisen Methoden nötige, Kreuzvalidierung der Ergebnisse anhand einer weiteren Stichprobe in vergleichbarer Größe im Rahmen dieser Studie aus organisatorischen Gründen nicht möglich gewesen.

Modell 1 und 2

Die beiden ersten Modelle, die im Rahmen dieser Analysen gerechnet wurden, basieren auf den bereits aus der Beantwortung der Forschungsfragen 1 bis 3 erhaltenen Informationen. Es wurden daher die Variablen *Praetest* und *Ist Tutor* in dieser Reihenfolge zur Vorhersage der Kriteriumsvariable *Posttest* in die Modelle aufgenommen. Mit diesen zwei Prädiktoren wurden die Modelle 1 und 2 (Tabelle 7.21) gerechnet.

[22] Die schrittweisen Methoden wählen zunächst den Prädiktor mit der höchsten Korrelation (Kriteriumsvalidität) aus. Falls die β-Gewichte dieses Prädiktors signifikant sind, wird der nächste Prädiktor hinzugenommen. Es ist dies jener mit den höchsten Semipartialkorrelationen. Beendet wird das Verfahren dann, wenn der letzte hinzugekommene Prädiktor nicht mehr signifikant ist.

Modell 3 und Modell 4

Danach wurden entsprechend der Stärke und der Signifikanz der Korrelationen, die Variablen *Muttersprache* (Modell 3) sowie *Geschlecht* und *Letzte Note in Physik* hinzugenommen (Modell 4). Die Hinzunahme der Variable *Muttersprache* schien plausibel, da sie in einer Situation von Bedeutung ist, in der es durch die Notwendigkeit zu erklären auf sprachlichen Ausdruck und sprachliche Fertigkeiten ankommt. Wie die Ergebnisse der Interessensforschung in den Naturwissenschaften zeigen (z.B. Häußler, et al., 1998) konnte auch von der Variable Geschlecht angenommen werden, dass sie Einfluss auf die Ergebnisse des Peer Tutoring Prozesses hat, sobald dieser sich mit den, bei Mädchen tendenziell unbeliebten, physikalischen Inhalten auseinandersetzt.

Modell		Regressions-koeffizient b	Standard-fehler	Standardi-sierte Beta β	T	Sig.
1	(Konstante)	3,808	0,430		8,862	0,000
	Praetest	0,456	0,082	0,426	5,565	0,000
2	(Konstante)	2,327	0,639		3,643	0,000
	Praetest	0,406	0,081	0,379	4,996	0,000
	Ist Tutor	1,956	0,639	0,232	3,063	0,003
3	(Konstante)	1,428	0,712		2,006	0,047
	Praetest	0,385	0,080	0,360	4,827	0,000
	Ist Tutor	2,215	0,633	0,263	3,499	0,001
	Muttersprache	1,155	0,437	0,195	2,643	0,009
4	(Konstante)	2,234	0,930		2,401	0,018
	Praetest	0,374	0,080	0,349	4,656	0,000
	Ist Tutor	2,367	0,644	0,281	3,676	0,000
	Muttersprache	0,836	0,502	0,141	1,666	0,098
	Geschlecht	-0,274	0,421	-0,048	-,650	0,517
	Note	-0,275	0,231	-0,103	-1,194	0,235

Tabelle 7.21: Standardisierte und nicht-standardisierte Regressionskoeffizienten zu vier MLR Modellen

Die Ergebnisse der MLR für alle vier Modelle sind in Tabelle 7.21 dargestellt. Die nicht-standardisierten Regressionskoeffizienten geben den absoluten Einfluss eines Prädiktors auf die Kriteriumsvariable *Posttest* an, sind aber abhängig von der Kodierung. Will man die Einflüsse verschiedener Variablen vergleichen, ist es sinnvoll die standardisierten Betas zu verwenden. Es sind dies die Regressionskoeffizienten der z-standardisierten

Variablen[23]. Verändert man die zugehörige Prädiktorvariable um eine Standardabweichung, ändert sich das Kriterium um Beta (β) Standardabweichungen.

Tabelle 7.22 zeigt die Korrelationen zwischen beobachteten und vorhergesagten Werten (R), den Standardschätzfehler, die Prüfgröße F und die Effektstärken. Wesentlich

Modell	R	R^2	Korrigiertes R^2	Standardfehler d. Schätzers	$F_{empirisch}$	Effektstärke f^2
1	0,426	0,181	0,175	2,554	30,970	0,22
2	0,483	0,233	0,222	2,481	21,104	0,30
3	0,520	0,270	0,254	2,429	17,004	0,37
4	0,529	0,280	0,253	2,430	10,574	0,39

Tabelle 7.22: Vier MLR-Modell im Vergleich: multiple Korrelationskoeffizienten (R) und Determinationskoeffizienten (R^2)

aufschlussreicher sind aber die um die Anzahl der Prädiktoren korrigierten Determinationskoeffizienten (korrigiertes R^2). Diese können als Verhältnis der Varianzen der vorhergesagten Werte zu den Varianzen der beobachteten Werte interpretiert werden.

R^2 drückt daher jenen Anteil der Varianz an der gesamten beobachteten Varianz aus, der durch das betreffende Modell erklärt werden kann. In diesem Sinne erreicht Modell 3 die größte aufgeklärte Varianz mit R^2 = 25,4 %. Modell 4 hat zwar mit R^2 = 25,3 % einen ähnlich hohen Anteil, arbeitet aber mit mehr Prädiktoren als Modell 3, von denen die letzten drei nicht signifikant sind.

Führt man darüber hinaus einen F-Test durch, der die Modellgüte beschreibt, indem er gegen die Nullhypothese testet, dass der vorhergesagte Anteil an der Kriteriumsvarianz Null ist, so sind alle hier angeführten Modelle höchstsignifikant mit einer Übertretungswahrscheinlichkeit von $p < 0{,}001$.

Zusätzlich wurden für die Beurteilung der durch die Intervention erzielten Effekte die Effektstärken berechnet. Für das Modell 3 beispielsweise ergibt sich eine Effektstärke

[23] Die z-Standardisierung entspricht der Transformation einer (μ;σ) normalverteilten Variable auf eine (0;1) normalverteilte Variable.

von $f^2 = 0{,}39$, was als eher starker Effekt interpretiert werden kann (Buehner & Ziegler, 2009).

Darüber hinaus wurden post hoc für die oben angeführten Modelle auch Kollinearitätsdiagnosen durchgeführt. Alle berechneten Konditionsindizes waren kleiner als 11,3, lagen somit unter der Grenze von 15, ab der man von mäßiger Kollinearität spricht (Bühner & Ziegler, 2009). Es ist daher nicht davon auszugehen, dass die Prädiktoren kollinear sind. Daher sind die standardisierten Beta-Gewichte ausreichend genau geschätzt und können problemlos interpretiert werden.

Alle hier durchgeführten ML Regressionen sind somit problemlos zu interpretieren, von der Charakteristik des Samples her, wie auch von den gewählten Prädiktoren.
Von den oben beschriebenen Modellen wird somit das Modell 3 präferiert. Alle Prädiktoren, die hier eingehen, sind zumindest auf einem Niveau von $p < 0{,}01$ signifikant. Das Modell zeigt zudem ein größtmögliches Maß an erklärter Varianz bei einer zufrieden stellenden Effektstärke und kommt gleichzeitig mit wenigen Prädiktorvariablen aus, die sich auch theoretisch gut erklären lassen.

7.2. Ergebnisse aus der Optik

Im zweiten Studienjahr, das zugleich den zweiten Teil der Datenerhebung darstellte, wurden, wie schon in Kapitel 4.3 ausführlich beschrieben, ausgewählte Kapitel aus der Optik behandelt. Das Forschungsdesign blieb im Vergleich zum ersten Studienjahr dasselbe: ein Praetest-Posttest-Follow-up-Test Design. Bis auf Klasse 1b waren auch die Schüler/innen in den Klassen dieselben wie im ersten Studienjahr mit dem Unterschied, dass die Schüler/innen um ein Jahr älter waren. Klasse 1a des ersten Studienjahres schied aus dem Grund aus, da es sich im vorangegangenen Schuljahr um eine 8. Schulstufe handelte, die in dieser Form im darauffolgenden Schuljahr nicht mehr existierte (Ende der Schulform mit der 8. Schulstufe). Anstelle dieser Klasse kam eine andere, ebenfalls eine 8. Schulstufe, aus derselben Schule hinzu.

Das Themengebiet der Optik war für alle Schüler/innen der Sekundarstufe 1 unbekannt. Für die 8. Schulstufe ist die Optik zwar im Lehrplan vorgeschrieben, jedoch wird sie erfahrungsgemäß erst am Ende dieser Schulstufe unterrichtet. Darüber hinaus wurden die Lehrkräfte gebeten, dass bis zum Abschluss des CAPT-Projektes dieses Thema im Regelunterricht nicht bearbeitet werden solle.

7.2.1. Analysen der Praetests und Auswahl der zu beforschenden Klassen

Wie in Kapitel 5 berichtet, wurde aufgrund der verwendeten Messinstrumente eine Einschränkung der zu beforschenden Schüler/innen auf die Sekundarstufe 1 getroffen. Daher scheint hier, wie schon in den Analysen zur Elektrizitätslehre, Klasse 5 nicht auf, da sie der Sekundarstufe 2 angehörte. Wie weiter unten ausführlich argumentiert, wird innerhalb der Sekundarstufe 1 auch die Klasse 6 in den Analysen nicht berücksichtigt werden.

Tabelle 7.23 gibt einen Überblick über die Klassen der Sekundarstufen, ihre Rolle im Tutoringprozess und die dabei behandelten Themen.

Klasse	Schulform[24]	Schulstufe	Rolle	Thema
1b	KMS	8	Tutoren	Schatten
2	NMS	7	Tutoren	Spiegel
3	NMS	7	Tutoren	Schatten
4	HS	8	Tutoren	Schatten
6	AHS-Unterstufe	8	Tutoren/Tutees	Spiegel/Schatten
7	NMS	7	Tutoren/Tutees	Schatten
8	NMS	7	Tutees	Spiegel
9	NMS	8	Tutoren	Schatten
10	NMS	8	Tutoren/Tutees	Spiegel

Tabelle 7.23: Übersicht über die beteiligten Klassen, die Rollen der Schüler/innen und die behandelten Themen.

Die Gründe, die zum Ausschluss der Klasse 6 aus den Analysen führten, lagen vor allem in der Sonderstellung, die diese Klasse innerhalb des Treatments innehatte: Nur in dieser Klasse wurden beide Themen aus der Optik behandelt, Schatten und Spiegel. Darüber hinaus erfuhr die Klasse eine wesentlich längere Intervention, da es zwei Mentorings, zu jedem der Themen eines, und zwei Tutorings gab, also insgesamt fünf Lerngelegenheiten. Zwar sollten nach den Ergebnissen aus dem vorangehenden Kapitel zur Elektrizitätslehre keine Dosiseffekte zu erwarten sein, allerdings wurden im Rahmen dieser Analysen lediglich der Unterschied zwischen zwei und drei Lerngelegenheiten untersucht. Dass es dabei keine signifikanten Unterschiede gibt lässt nicht den Schluss zu, dass es im Vergleich mit fünf Lerngelegenheiten, wie es in diesem Setting für Klasse 6 das Fall gewesen ist, ebenfalls keine Unterschiede geben würde. Der Einfluss einer derartigen Fülle von Lerngelegenheiten wäre separat zu untersuchen, was aber den Rahmen dieser Arbeit sprengt und a priori auch nicht Gegenstand der Forschungsfragen war. Daher wird dieser Frage in weiterer Folge nicht nachgegangen.

[24] AHS ... Allgemein bildende höhere Schule; HS ... Hauptschule; KMS ... Kooperative Mittelschule ; NMS ... Neue Mittelschule

Darüber hinaus wich Klasse 6 im Praetest vom restlichen Sample in einem weiteren Punkt stark ab. Es handelte sich hier um die einzige AHS-Klasse innerhalb der Sekundarstufe 1. Alle anderen Klassen dieser Altersstufe waren Formen der Hauptschule[25]. Es wurden zwar keine kognitiven Parameter erhoben, aber die hochsignifikant besseren Ergebnisse im Praetest lassen auf eine eventuelle höhere Leistungsfähigkeit der AHS-Schüler/innen im Vergleich zu Hautschüler/innen schließen. Somit stellte sich die Interpretation eines Vergleichs der Posttest-Ergebnisse dieser Klasse mit allen anderen Gruppen als problematisch dar.

Ergebnisse der detaillierten Analysen der Wissens-Praetests zum Thema Spiegel

Tabelle 7.24 zeigt die Mittelwerte der Klassen in diesem Wissens-Praetest zum Thema Spiegel.

Klasse	Mittelwert	SD	N
2	1,78	1,68	23
6	5,70	1,84	20
8	3,10	1,91	10
10	3,28	1,74	18
Gesamt	3,45	2,32	71

Tabelle 7.24: Mittelwerte der Praetests zum Thema Spiegel

Auf den ersten Blick fallen Klasse 2 (besonders wenige Punkte) und Klasse 6 (besonders viele Punkte) auf. Um diesen Eindruck statistisch zu untermauern, wurden Kontrasttests durchgeführt.

Kontrast-Koeffizienten				
Kontrast	Klasse			
	2	6	8	10
1	1	-3	1	1
2	-3	1	1	1

Tabelle 7.25: Überblick über die durchgeführten Vergleiche: Kontrast 1 vergleicht Klasse 6 mit allen anderen, Kontrast 2 Klasse 2 mit allen anderen.

[25] Unter dem Begriff *Hauptschule* werden in dieser Arbeit verschiedene Hauptschulformen zusammengefasst wie: Neue Mittelschule (NMS), Kooperative Mittelschule (KMS) oder Hauptschulen mit einem speziellen Schwerpunkt.

Kontrast-Test				
Kontrast	Kontrastwert	T	Df	Signifikanz (2-seitig)
1	8,94	-6,06	34,1	0,000
2	6,73	5,01	44,5	0,000

Tabelle 7.26: Ergebnisse der Kontrast-Tests (keine Gleichheit der Varianzen angenommen) zum Thema Spiegel

Aus Tabelle 7.25 geht hervor, dass durch Kontrast 1 die Praetest-Ergebnisse der Klasse 6 mit allen anderen Klassen verglichen wurden. Mit Kontrast 2 wurde ein Vergleich der Klasse 2 mit allen anderen Klassen durchgeführt. Die Ergebnisse der Kontrast-Tests sind in Tabelle 7.26 zu sehen. Beide Tests zeigen ein höchstsignifikantes Ergebnis ($p < 0,001$). Das bedeutet, dass Klasse 6 höchstsignifikant über dem Durchschnitt abschneidet, während Klasse 2 höchst signifikant darunter ist. Bei diesen Kontrast-Tests wurde für die Analysen die strengere Annahme *„Varianzen nicht gleich"* gewählt, obwohl der Test auf Homogenität der Varianzen (Levene-Test) nicht signifikant war. Das ist eine mittlerweile verbreitete Vorgangsweise, weil dieser weitere Test einen zusätzlichen statistischen Fehler verursachen könnte. Prinzipiell wäre es aufgrund dieses Kontrast-Tests möglich beide Klassen (6 und 2) von der weiter gehenden Analyse auszuschließen. Die statistischen Parameter sollen aber nicht die alleinige Entscheidungsbasis darstellen, sondern vielmehr die folgenden inhaltlichen Überlegungen. Bedenkt man, dass Klasse 6 ein stark abweichendes Treatment erfahren hat, so ist dies ein wichtigerer, weil inhaltlich begründeter und nicht bloß statistische ausgewiesener, Grund um sie auszuschließen, im Unterschied zu Klasse 2. Darüber hinaus zeigte sich, dass ein Ausschluss von Klasse 6 aus dem Vergleich dazu führt, dass Klasse 2 nicht mehr hochsignifikant schlechter als die anderen Klassen abschneidet. Post-hoc bildet sie im Tukey-HSD Test auf Basis von $\alpha = 0,05$ eine homogene Gruppe mit den verbleibenden Klassen (8 und 10). Aus diesen Gründen wurde Klasse 2 in die weiteren Analysen eingebunden, Klasse 6 jedoch nicht.

Ergebnisse der detaillierten Analysen der Wissens-Praetests zum Thema Schatten

Führt man vergleichbare Analysen zum zweiten Thema dieses Studienteils, dem Schatten durch, ergibt sich ein ähnliches Bild. Vergleicht man in Tabelle 7.27 die Ergebnisse der einzelnen Klassen im Wissens-Praetest, so zeigt sich, dass Klasse 6 wiederum deutlich besser abschneidet, während eine der Klassen, in diesem Fall

Klasse 3, deutlich schlechter abschneidet. Ein Kontrast-Test (Tabelle 7.28) untermauert diese Behauptung statistisch, wobei hier Kontrast 1 Klasse 6 gegen alle anderen testet und Kontrast 2 Klasse 3 gegen alle anderen testet.

Klasse	N	Mittelwert	SD
1b	14	5,50	1,74
3	15	3,07	1,87
4	24	6,21	1,99
6	20	8,25	1,86
7	18	5,39	2,45
9	18	5,56	2,01
Gesamt	109	5,82	2,47

Tabelle 7.27: Mittelwerte der Praetests zum Thema Schatten

Kontrast-Test				
Kontrast	Kontrastwert	T	Df	Signifikanz (2-seitig)
1	15,57	5,91	19,7	0,000
2	-15,53	-6,62	30,2	0,000

Tabelle 7.28: Ergebnisse der Kontrast-Tests (keine Gleichheit der Varianzen angenommen) zum Thema Schatten

Leider bleibt in diesem Fall auch bei Ausschluss der Klasse 6 der Unterschied der Klasse 3 zu den verbleibenden Klassen hochsignifikant erhalten. Dennoch wurde diese Klasse nicht von der weiteren Analyse ausgeschlossen, da sie inhaltlich ein vergleichbares Treatment wie die anderen Klassen erhalten hat. Für die weitere Vorgangsweise und vor allem die Interpretation der Daten wird dieser Unterschied zu den restlichen Klassen aber von Bedeutung sein. Auch gilt es mögliche Ursachen für das schlechtere Abschneiden von Klasse 3 festzumachen und die wissensmäßige Entwicklung genau zu verfolgen.

Die Analysen zu den Praetests zeigen somit, dass man bei einem Vergleich der Posttest-Ergebnisse mit den Praetest-Ergebnissen achtsam sein muss und auftretende Unterschiede nicht leicht interpretierbar sind. Außerdem ist die nicht geringe Zahl an

fehlenden Fragebögen (13,6 %) zu beachten. Gleichwohl wurde zur ersten Orientierung ein Gesamtvergleich über das gesamte Sample hinweg angestellt.

7.2.2. Beschreibung des Samples

Im zweiten Jahr der Untersuchung änderten sich die statistischen Eckdaten des Samples ein wenig, da die Klasse 1b physisch nicht mit der Klasse 1a aus dem ersten Studienjahr identisch war und Klasse 6, wie oben beschrieben, bei diesen Analysen nicht berücksichtigt wurde. Tabelle 7.29 und Abbildung 7.13 geben einen Überblick über alle N = 141 Schüler/innen, die im zweiten Studienjahr an dieser Untersuchung beteiligt waren und deren Daten ausgewertet wurden.

Klasse	Schulstufe	Schülerzahl
Klasse 1b	8	14
Klasse 2	7	23
Klasse 3	7	15
Klasse 4	8	24
Klasse 7	7	18
Klasse 8	7	11
Klasse 9	8	18
Klasse 10	8	18
Gesamt		**141**

Tabelle 7.29: Klassen und Anzahl der Schüler/innen

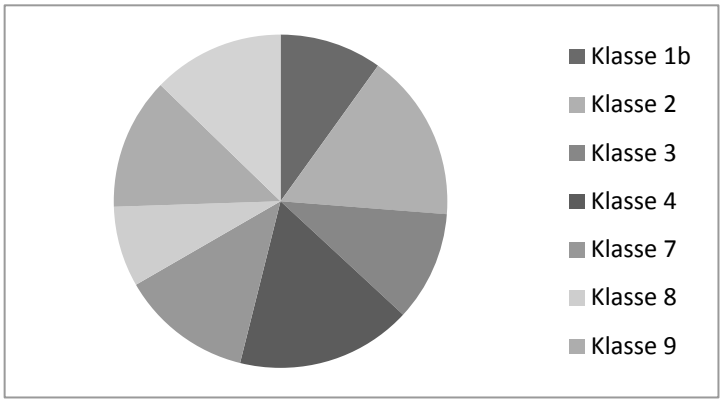


Abbildung 7.13: Klassen und Anzahl der Schüler/innen

Die Verteilung der männlichen und weiblichen Schüler/innen (59 % und 41 %) ist der in Kapitel 7.1.1 beschriebenen sehr ähnlich. Der Unterschied zwischen beiden Samples besteht darin, dass es sich hier ausschließlich um Hauptschulklassen handelt. Die Ursache für den Überhang der männlichen Schüler liegt wiederum im Schultyp begründet (Statistik-Austria, 2013). Abweichungen der Prozentsätze, die im Sample zu finden sind, zu den Daten der Statistik Austria sind im t-Test auf Basis einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ nicht signifikant.

Der Anteil an Schüler/innen mit deutscher Muttersprache liegt bei 73 %, jener mit einer anderen Muttersprache bei 27 %. Eine klassenweise Analyse ergab, dass in Klasse 2, mit 80 %, und in Klasse 3, mit 75 %, ein besonders hoher Anteil an Schüler/innen nicht-

deutscher Muttersprache zu finden war (Abbildung 7.14). Das entspricht hinsichtlich der Muttersprachen dem Bild, das sich in Ballungsräumen wie Wien und Umgebung auch in anderen öffentlichen Hauptschulen bietet. Klasse 2 und 3 sind aus eben einer solchen öffentlichen Schule, während die Klassen 1b, 7, 8, 9 und 10 Privatschulen sind und Klasse 4 eine Praxisschule einer Ausbildungseinrichtung für Lehrer/innen ist.

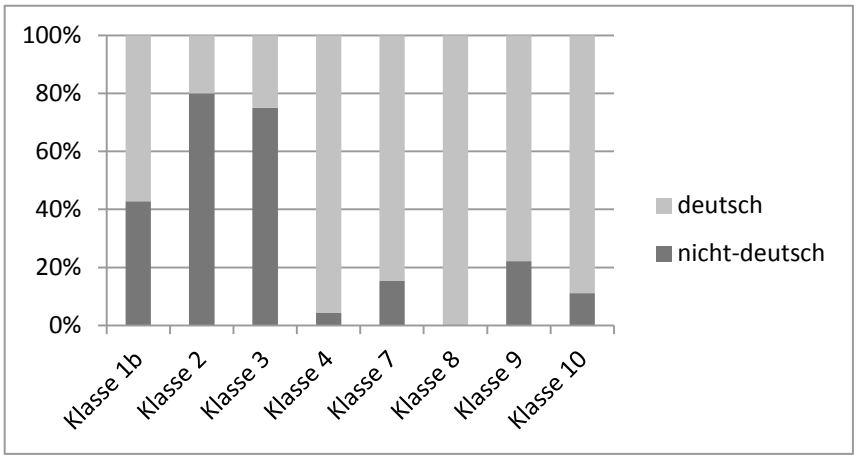


Abbildung 7.14: Anteile der Schüler/innen mit deutscher bzw. einer anderen Muttersprache nach Klasse dargestellt

Aus Abbildung 7.15 geht hervor, wie häufig die drei möglichen Rollen, die Schüler/innen innerhalb des Tutoringprozesses einnehmen konnten, vorkommen. *Tutoren* waren jene,

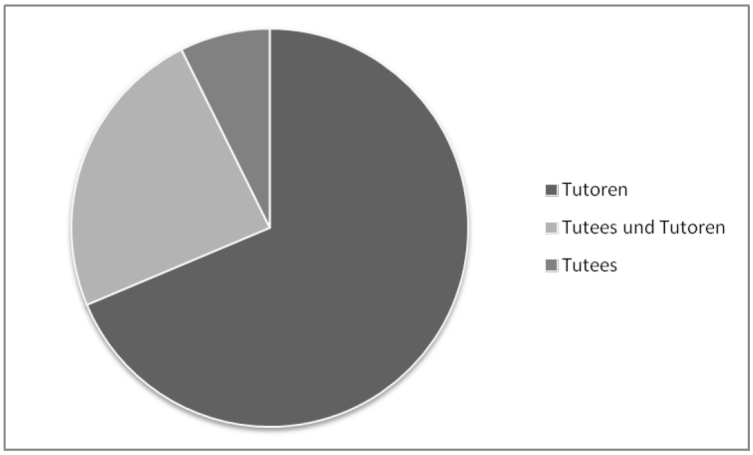


Abbildung 7.15: Verteilung der Rollen innerhalb des Tutoringprozesses

die ausschließlich die aktive Rolle einnahmen, als *Tutees und Tutoren* wurden jene bezeichnet, die einmal die passive Rolle der Tutees inne hatten und ein zweites Mal in

der aktiven Rolle als Tutoren tätig waren, während die *Tutees* nur in der passiven Tutee-Rolle waren.

Die Schüler/innen in den drei möglichen Rollen teilten sich zusätzlich auf die Themen Schatten und Spiegel auf (vgl. Abbildung 7.16). Die untersuchte Stichprobe besteht daher aus N_{Sp} = 46 Probanden, für die die Wissenstests zum Thema Spiegel analysiert wurden und N_{Sch} = 75 Probanden, für die die Wissenstests zum Thema Schatten ausgewertet wurden.

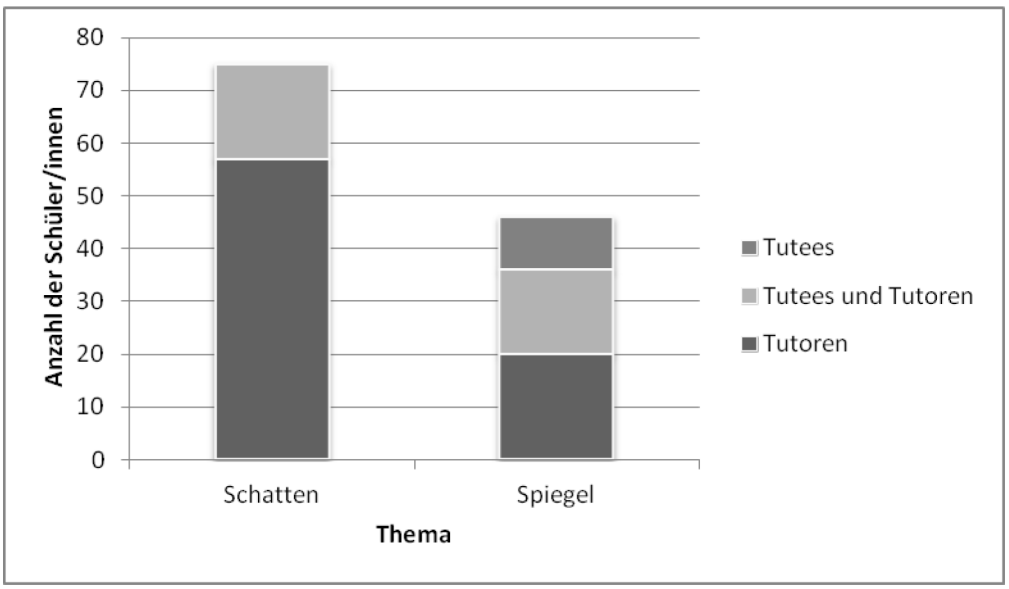


Abbildung 7.16: Übersicht über die Rollen- und Themenverteilung (ohne Klasse 6)

Die Aufteilung des Samples nach Rollen *und* Themen bedingt für die einzelnen Gruppen sehr kleine Stichprobenumfänge. Bei der Wahl der statistischen Methoden wird darauf besonders eingegangen. Insbesondere sind dieselben Tests und Vergleiche wie in Kapitel 7.1 zur Elektrizitätslehre nicht anwendbar.

7.2.3. Vergleich der Praetest-Ergebnisse mit den Posttest-Ergebnissen – Optik

Im vorangegangenen Kapitel wurden die Voraussetzungen, die die Schüler/innen mitbrachten, analysiert. Auf Basis dessen werden im Rahmen dieses Kapitels die Forschungsfragen 1 und 6 beantwortet. Zur besseren Orientierung werden diese Fragen wieder am Beginn des entsprechenden Abschnittes wiederholt.

***Forschungsfrage 1**: Weisen Schüler/innen der Sekundarstufe 1 nach der CAPT-Intervention bessere kognitive Testergebnisse auf und wenn, wie stark ist der gemessene Effekt?*

Um einen ersten Überblick über das Sample und die Wirksamkeit der Intervention zu erhalten, wurden, getrennt nach den Themen Schatten und Spiegel, t-Tests für abhängige Stichproben durchgeführt. Diese Tests sollen zeigen, ob sich die Mittelwerte der Praetests signifikant von denen der Posttests unterschieden, wobei durch die Intervention von einer Verbesserung und nicht von einer Verschlechterung der Ergebnisse ausgegangen wird. Es handelt sich somit um eine gerichtete Fragestellung, es wird einseitig getestet mit der Nullhypothese, dass keine Verbesserungen in der Population durch die Intervention zu erwarten sind. Die Alternativhypothese ist in diesem Fall die gerichtete Hypothese, dass die Schüler/innen im Posttest besser abschneiden als im Praetest. Zuvor wurde überprüft, ob die Voraussetzungen für die Durchführung von t-Tests gegeben sind.

Prüfen der Voraussetzungen zur Durchführung von t-Tests

Um t-Tests durchführen zu können bzw. deren Ergebnisse interpretieren zu können, müssen drei Voraussetzungen erfüllt sein (Buehner & Ziegler, 2009):

- Die Messwerte der Personen müssen unabhängig sein.
- Die Messwerte müssen intervallskaliert sein.
- Die Differenzen der getesteten Variablen müssen in der Grundgesamtheit und in der Stichprobe normal verteilt sein.

Die Erfüllung der ersten Voraussetzung bedeutet, dass sich die Schüler/innen bei den Wissenstests nicht gegenseitig beeinflussen durften. In anderen Worten: Sie durften nicht voneinander abschreiben. Diese Voraussetzung lässt sich zwar nicht wirklich testen, war aber durch die Aufsicht bei jedem Test bestmöglich erfüllt worden.

Um die zweite Voraussetzung zu erfüllen, prüft man, ob die Messwerte intervallskaliert sind. Für die Wissenstests, entsprechend der Richtigkeit der Antworten, wurden Punkte vergeben (siehe Kapitel 5.2). Das bedeutet, dass denjenigen Schüler/innen, die mehr wussten, mehr Punkte zugeordnet wurden. Für eine Intervallskala muss darüber hinaus die Differenz dieser Punkte sinnvoll interpretierbar sein: Ist eine Differenz im Wissen

zwischen zwei Personen vorhanden, muss sich diese Differenz auch in den vergebenen Punkten widerspiegeln. Hinzu kommt, dass eine Differenz von z.B. 1,5 Punkten immer das Gleiche aussagen soll. Auch das kann im Rahmen dieses Wissenstests als erfüllt angesehen werden, da aus den vergebenen Punkten ein Summenwert gebildet wurde.

Um die dritte Voraussetzung zu testen, wurde die Normalverteilungsannahme im Sample beurteilt. Prinzipiell ist das mittels Kolmogoroff-Smirnoff-Test oder über die Beurteilung der Histogramme möglich. Es wurde hier auf den Kolmogoroff-Smirnoff-Test verzichtet, da dieser als nonparametrischer Test zwar sehr stabil ist, aber eine geringe Teststärke aufweist. Stattdessen wurden Histogramme beurteilt. Dafür wurden für die Variablen *Spiegel-gain* (Differenz zwischen Posttestergebnis und Praetestergebnis zum Thema Spiegel) und der Variable *Schatten-gain* (Differenz Posttestergebnis – Praetestergebnis zum Thema Schatten) Histogramme erstellt (Abbildung 7.17) und beurteilt.

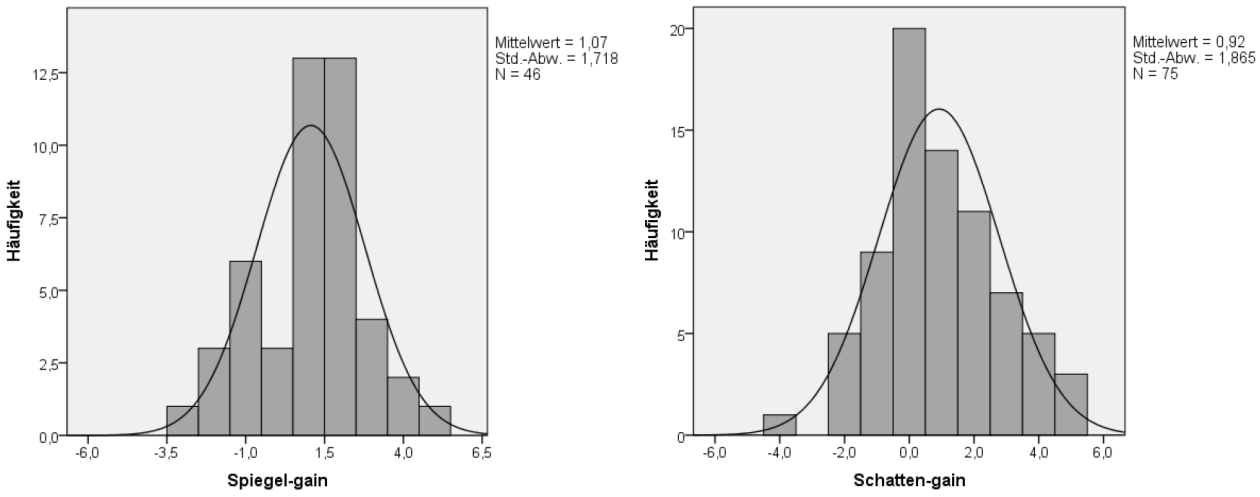


Abbildung 7.17: Histogramme der Variablen *Spiegel-gain* und *Schatten-gain* (Posttest – Praetest)

Die Beurteilung der Histogramme liefert eine zufrieden stellende Übereinstimmung mit der Normalverteilung (schwarze Linie), sodass die dritte Bedingung, gegen deren Verletzung der t-Test allerdings robust wäre, nicht nur in der Stichprobe, sondern auch in der Population als erfüllt angesehen werden kann. T-Tests können somit in diesem Sample und für diese Variablen ohne Bedenken durchgeführt werden.

Was die Behandlung der fehlenden Werte betrifft, wurde mit *listwise deletion* gearbeitet (vgl. 4.5). Für die Tests zum Thema Spiegel ist das in Anbetracht der fehlenden Fragebögen zu beiden Testzeitpunkten eine korrekte Vorgangsweise. Für die Test zum Thema Schatten liegt man jedoch schon im oberen Bereich dessen, was noch toleriert werden kann. Aus Gründen der Vergleichbarkeit der Daten wurde aber für alle Themen dieselbe Vorgangsweise gewählt. Es ist jedoch zu beachten, dass die Populationsparameter ein wenig verzerrt geschätzt werden könnten.

Eigentliche t-Tests für die Themen Spiegel und Schatten

Die beiden folgenden Tabellen geben einen Überblick über die Stichprobengröße, die Mittelwertunterschiede (Tabelle 7.30) und die Korrelationen (Tabelle 7.31) der

Statistik bei gepaarten Stichproben			
	Mittelwert	N	SD
Spiegel-post	3,65	46	2,17
Spiegel-prae	2,59	46	1,88
Schatten-post	6,37	75	2,34
Schatten-prae	5,45	75	2,29

Tabelle 7.30: Mittelwerte in Prae- und Posttests im Vergleich

getesteten Variablen. Es ist zu erkennen, dass für jedes der Themen der Mittelwert des Posttests höher ist als der des Praetests, die Schüler/innen also im Wissenstest, der nach dem Treatment durchgeführt wurde, besser abgeschnitten haben.

Korrelationen bei gepaarten Stichproben				
		N	Korrelation	SD
Spiegel-post	**&**	46	0,649	0,000
Spiegel-prae				
Schatten-post	**&**	75	0,675	0,000
Schatten-prae				

Tabelle 7.31: Korrelationen zwischen Prae- und Posttests

Die Mittelwertunterschiede und deren Signifikanzen sind in Tabelle 7.32 angeführt. Es zeigt sich, dass für jedes der Themen der Mittelwertunterschied zwischen Praetest und Posttest etwa einen Punkt beträgt und dass diese Unterschiede höchstsignifikant sind ($p < 0{,}001$).

Test bei gepaarten Stichproben						
		Gepaarte Differenzen				
		Mittelwert	SD	T	df	Sig. (2-seitig)
Paar 1	Spiegel-post – Spiegel-prae	**1,065**	1,718	4,205	45	**0,000**
Paar 2	Schatten-post – Schatten-prae	**0,920**	1,866	4,271	74	**0,000**

Tabelle 7.32: t-Test auf Mittelwertunterschiede zwischen Praetests und Posttests

Zusätzlich wurden post-hoc die Teststärke und die Effektstärke nach Cohen (1988) berechnet, um eine fundierte Interpretation des Effektes des Treatments zu vornehmen und seine praktische Bedeutung einschätzen zu können.

Thema	Effektstärke	Teststärke[26]
Spiegel	0,62	ca. 1
Schatten	0,49	ca. 1

Tabelle 7.33: Effektstärken nach Cohen und Teststärken der t-Tests aus Tabelle 7.32

In diesem Fall, da die Teststärke fast 1 ist, ist es sehr wahrscheinlich, dass das Treatment wirklich zu einer Verbesserung des Wissens in den getesteten Bereichen führt.

Nachdem mit diesen Analysen ein grober Überblick über die Situation geschaffen wurde, folgen nun tiefer gehende Analysen auf Klassenebene, die aufgrund der sehr unterschiedlichen Praetests im Wissensbereich notwendig erscheinen.

[26] Das verwendete Programm, G*Power 3.1, gibt 7 Nachkommastellen an, in diesem Fall 9er.

7.2.4. Klassenweise Vergleiche der Wissenstests – Schatten

Dieses Kapitel widmet sich der Beantwortung der

Forschungsfrage 6: *Welche Einflüsse auf die Posttest-Ergebnisse ergeben sich aufgrund der unterschiedlichen Klassenzugehörigkeiten?*

Da sich Durchführung und Voraussetzungen der Schüler/innen im Themenbereich Optik stark von denen im Themenbereich Elektrizitätslehre unterscheiden, ist es prinzipiell nötig, einen anderen, adäquateren Zugang zu den statistischen Auswertungen zu wählen.

Bei der Durchführung ist zu bedenken, dass hier zwei verschiedene Inhaltsbereiche, nämlich Schatten und Spiegel bearbeitet wurden. Da auch die Voraussetzungen (Praetests), die die einzelnen Klassen mitbrachten und die im vorangegangenen Kapitel ausführlich beschrieben wurden, sehr unterschiedlich waren, ist es sinnvoll in weiterer Folge die Ergebnisse auf Klassenebene zu untersuchen und zu identifizierende Haupteffekte und deren Wechselwirkungen zu analysieren. Es werden daher Interaktionsanalysen angestellt.

Als Haupteffekte sind der Testzeitpunkt (Praetest – Posttest) und die Klassenzugehörigkeit von Interesse. Während der Messwiederholungsfaktor zweistufig ist, ist der Gruppenfaktor mehrstufig. Es liegt somit ein zweifaktorielles Design vor, bei dem allerdings Innersubjekteffekte (Messwiederholung) und Zwischensubjekteffekte (Klassenzugehörigkeit) gemischt werden. Daher spricht man innerhalb eines Allgemeinen Linearen Modells (ALM) auch von einer zweifaktoriellen Varianzanalyse mit Messwiederholung (gemischtes Design).

Die Fragestellungen, die damit beantwortet werden können, sind einerseits, ob das Treatment einen Effekt zeigt (Vergleich der Testzeitpunkte) und andererseits, ob dieser Effekt von der Klassenzugehörigkeit abhängt. Anders ausgedrückt werden parallel zu individuellen Vergleichen klassenweise Vergleiche angestellt.

Die getesteten Nullhypothesen lauten:

- H1 – Haupteffekt Klasse: Die Gruppenmittelwerte der verschiedenen Klassen unterscheiden sich nicht und gehören daher einer Grundgesamtheit an. Das

bedeutet, dass die Gruppenbedingungen keinen Einfluss auf die Testergebnisse im Wissenstest haben.

- H2 – Haupteffekt Testzeitpunkt: Die Mittelwerte der verschiedenen Messzeitpunkte unterscheiden sich nicht. Das bedeutet, dass der Messzeitpunkt (Praetest und Posttest) keinen Einfluss auf das Testergebnis hat.
- H 3 – Wechselwirkungen: Es gibt keine wechselseitige Beeinflussung der beiden Faktoren Testzeitpunkt und Klasse. Das bedeutet, dass sich die individuellen Mittelwerte linear aus denen der Klassen und des Testzeitpunktes ergeben.

Die zugehörigen Alternativhypothesen lauten:

- A1 – Die Gruppenmittelwerte der verschiedenen Klassen unterscheiden sich (ungerichtet). Die Klassenzugehörigkeit hat somit Einfluss auf das Testergebnis. Die Klassen gehören somit verschiedenen Grundgesamtheiten an.
- A2 – Der Mittelwert zum Testzeitpunkt 2 (Posttest) ist größer als der zum Testzeitpunkt 1 (Praetest). Das bedeutet, dass die Schüler/innen durch das Treatment ihr Wissen verbessern (gerichtet).
- A3 – Die Mittelwerte setzen sich nicht additiv aus den beiden Haupteffekten Klasse und Testzeitpunkt zusammen. Es gibt ein Zusammenwirken der beiden Faktoren.

Die Hypothesentests zu H1 und H2 wurden mittels einer zweifaktoriellen Varianzanalyse mit Messwiederholung (gemischtes Design) durchgeführt. Eine Bewertung der Wechselwirkungen (H3) über Interaktionsdiagramme soll zeigen, ob die Haupteffekte interpretierbar sind und wenn ja, in welche Richtung.

Prüfen der Voraussetzungen

Vor der Durchführung der Hypothesentests wurden die Voraussetzungen, unter denen eine derartige Varianzanalyse durchführen darf bzw. deren Ergebnisse sinnvoll interpretierbar sind, genau geprüft (Buehner & Ziegler, 2009). Es ergibt sich die folgende Liste an Voraussetzungen:

- Intervallskalenniveau der anhängigen Variablen
- Normalverteilung der Messwerte in allen Teilstichproben
- Homogenität der Gruppenvarianzen

- Homogenität der Varianzen und Kovarianzen zu den unterschiedlichen Testzeitpunkten (Sphärizität)
- Balanciertheit des Designs

Ähnlich wie in Kapitel 7.1.3 bereits argumentiert, liegt hier eine Intervallskalierung der abhängigen Variablen, also der Posttest-Ergebnisse, vor. Die Annahme der Normalverteilung in allen Teilstichproben wurde mittels Histogrammen geprüft und als ausreichend interpretiert.

Die Homogenität der Gruppenvarianzen wurde mit einem Levene-Test geprüft. Im aktuellen Fall sind die Gruppenvarianzen vergleichbar, da der Levene-Test nicht signifikant wurde[27].

Da im hier berichteten Fall lediglich zwei Testzeitpunkte analysiert wurden, entfällt die Prüfung der Sphärizitätsannahme[28].

Die Balanciertheit des Designs liegt vor, wenn zu jeder Testperson ein Messwert für jeden Testzeitpunkt vorliegt. Die Erfüllung dieser Voraussetzung wurde sichergestellt, indem jene Fälle aus der Analyse ausgeschlossen wurden, für die entweder der Prae- oder der Posttest nicht vorlag. Das führte zu einer Verkleinerung des Samples, da es immer wieder Schüler/innen gab, die bei einem der beiden Testzeitpunkte fehlten.

Tabelle 7.34 bietet im Überblick die klassenweisen Mittelwerte zu den Prae- und Posttests zum Thema Schatten. Für jede Klasse ist eine Erhöhung des Mittelwertes im Posttest gegenüber dem Praetest zu erkennen. Die relativ große Standardabweichung erlaubt eine Beurteilung der Differenzen erst anhand von elaborierteren Tests.

[27] Falls er signifikant gewesen wäre, ist ein F_{max} – Test durchzuführen (vgl. S. 134).

[28] Die Spärizitätsannahme soll sicher stellen, dass die Varianzen und Kovarianzen zwischen den Messzeitpunkten gleich sind und macht daher erst ab drei Messzeitpunkten Sinn. Allenfalls würde das mit dem Mauchly-Test getestet werden. Bei Verletzung der Sphärizitätsannahme gibt es statistische Möglichkeiten der Korrektur (z.B. Greenhouse-Geissler-Korrektur).

Deskriptive Statistiken				
	Klasse	Mittelwert	SD	N
Schatten-Prae	1b	5,91	1,70	11
	3	3,18	2,14	11
	4	6,22	2,04	23
	7	5,39	2,45	18
	9	5,75	2,01	12
	Gesamt	5,45	2,29	75
Schatten-Post	1b	6,27	1,85	11
	3	3,73	2,05	11
	4	7,57	2,02	23
	7	6,56	1,76	18
	9	6,33	2,64	12
	Gesamt	6,37	2,34	75

Tabelle 7.34: Klassenweise Vergleiche der Prae- und Posttests, Thema Schatten

Tests der Innersubjekteffekte					
Quelle	df	F	Sig.	Partielles Eta-Quadrat	Beobachtete Schärfe[a]
Testzeitpunkt	1	12,549	**0,001**	0,152	0,937
Testzeitpunkt * Klasse	4	0,826	**0,513**	0,045	0,251
Fehler (Testzeitpunkt)	70				
a. Unter Verwendung von Alpha = ,05 berechnet					

Tabelle 7.35: Effekte des Testzeitpunktes und Wechselwirkungen Testzeitpunkt-Klasse; Thema Schatten

Die Ergebnisse der Testung der Hypothesen H2 und H3 sind in Tabelle 7.35 dargestellt. Es zeigt sich, dass der Testzeitpunkt (H2) höchstsignifikanten Einfluss hat, daher im berichteten Fall die Schüler/innen ein höchstsignifikant besseres Ergebnis im Posttest haben. Die Wechselwirkung Testzeitpunkt und Klasse (H3) ist nicht signifikant ($p = 0{,}513$), jedoch könnte die Effektstärke (partielles $\eta^2 = 0{,}045$), trotz einer geringen

Teststärke (beobachtete Schärfe = 0,251), auf eine praktische Bedeutung des Effektes hinweisen. Das kann aus dem Grund passieren, da hier die Zellengröße sehr klein ist. Es würde für die Praxis bedeuten, dass die Klassenzugehörigkeit sehr wohl einen nicht unbedeutenden Einfluss auf das Testergebnis hat, wenn auch keinen signifikanten.

In Tabelle 7.36 folgt eine Darstellung der Zwischensubjekteffekte, daher es wird der Haupteffekt Klassenzugehörigkeit, also ein möglicher klassenweiser Unterschied getestet (H1). Der Levene-Test auf Gleichheit der Fehlervarianzen ist nicht signifikant, daher müssen keine Korrekturen berücksichtigt werden. Der Haupteffekt Klasse ist hier höchstsignifikant, daher unterscheiden sich die Klassen höchstsignifikant in ihrem Wissenszuwachs. Jedoch ist die Effektstärke dieses Effekts (hier: Partielles Eta-Quadrat) mit 0,27 moderat, die Teststärke[29] allerdings sehr hoch (0,99).

Tests der Zwischensubjekteffekte

Quelle	df	F	Sig.	Partielles Eta-Quadrat	Beobachtete Schärfe[a]
Konstanter Term	1	641,365	0,000	0,902	1,000
Klasse	4	6,428	**0,000**	0,269	0,986
Fehler	70				

a. Unter Verwendung von Alpha = ,05 berechnet

Tabelle 7.36: Klassenmittelwerte werden verglichen und auf ihre Signifikanz geprüft; Thema Schatten

Um sicher zu gehen, ob diese beiden Haupteffekte interpretierbar sind, werden die Wechselwirkungsdiagramme der beiden Effekte untersucht. Eine grafische Darstellung des Haupteffektes Testzeitpunkt liefert Abbildung 7.19, eine für die Klassenzugehörigkeit Abbildung 7.18[30].

[29] Die Teststärke wurde hier mittels SPSS 18 berechnet und wird dort als Beobachtete Schärfe ausgewiesen. Eine Berechnung mittels G*Power 3.1 würde bessere Werte liefern, da ein anderer Algorithmus verwendet wird. Im Falle dieser großen Effektstärken wurde aber darauf verzichtet.

[30] In den Abbildungen zu den Wechselwirkungsdiagrammen wurden hier, wie auch im nächsten Kapitel, ausschließlich aus Gründen der Übersichtlichkeit die Verbindungslinien beibehalten, obwohl sie bei diskreten Werten keine Bedeutung haben.

Auf Basis dieser Diagramme lassen sich die beiden Haupteffekte folgender Maßen interpretieren: In Abbildung 7.19 ist zu erkennen, dass alle Klassen im Posttest einen Auf Basis dieser Diagramme lassen sich die beiden Haupteffekte folgender Maßen

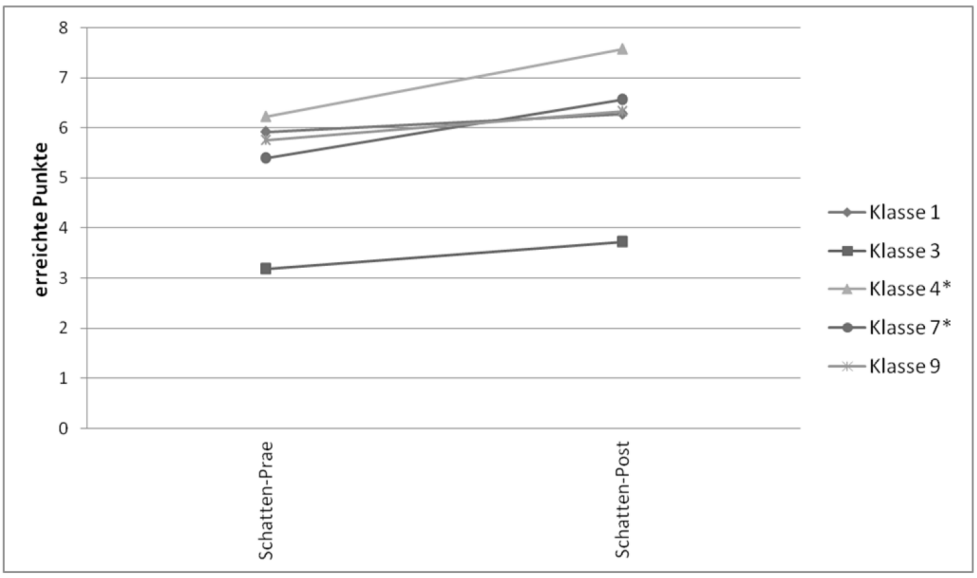


Abbildung 7.19: Darstellung des Haupteffektes Testzeitpunkt für die einzelnen Klassen; Thema Schatten

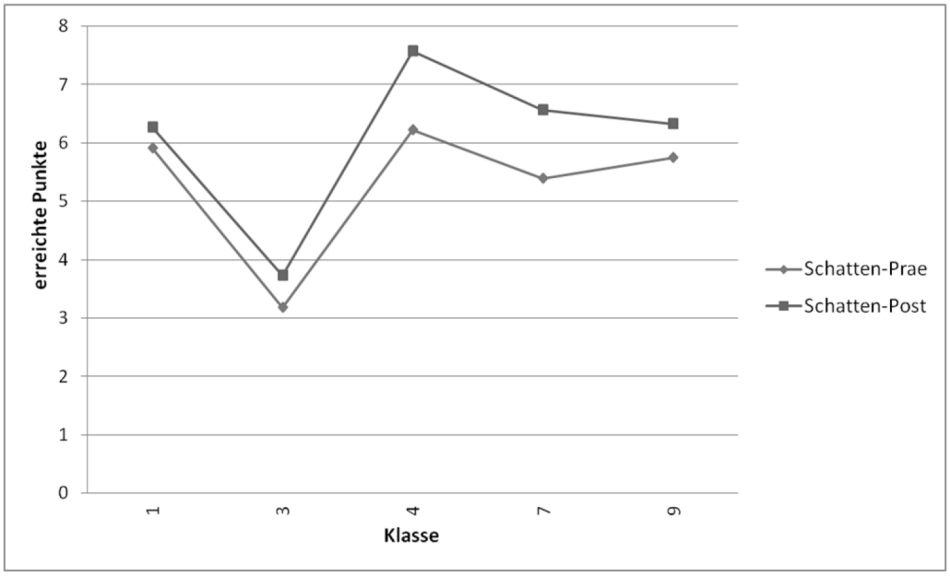


Abbildung 7.18: Darstellung des Haupteffektes Klasse; Thema: Schatten

interpretieren: In Abbildung 7.19 ist zu erkennen, dass alle Klassen im Posttest einen höheren Punktescore erreichen als im Praetest. Für jede der Klasse ist somit die Rangordnung erhalten, nämlich dass der Posttest einen niedrigeren Rang hat als der Praetest (höhere Punktezahl bedeutet niedrigeren Rang; vgl. Notenskala). Der

Haupteffekt Testzeitpunkt ist somit auf Basis des Signifikanzniveaus (Tabelle 7.35) und der Interaktionsdiagramme eindeutig dahingehend interpretierbar, dass Schüler/innen auf Basis der CAPT-Interventionen einen Wissenszuwachs zeigen.

Abbildung 7.18 lässt erkennen, dass hingegen die Rangordnung der Klassen nicht erhalten bleibt. So liegt z.B. Klasse 1 im Praetest an zweiter Stelle, im Posttest nur mehr an vierter Stelle. Das bedeutet jedoch, dass der Haupteffekt Klasse nicht eindeutig interpretierbar ist. Die Klassenzugehörigkeit lässt somit, obwohl der Effekt hochsignifikant ist, keine klaren Aussagen darüber zu, *wie* eine Klasse im Posttest abschneiden wird, lediglich darüber *dass* sie besser abschneiden wird als im Praetest (Bühner & Ziegler, 2009).

7.2.5. Klassenweise Vergleiche der Wissenstests – Spiegel

Im Folgenden sind äquivalente Analysen wie auf den vorangehenden Seiten zum zweiten Thema innerhalb der Optik, dem Spiegel, durchgeführt worden. Die Prüfung der Voraussetzungen erfolgte wie im vorigen Kapitel beschrieben. Es wurden die gleichen Hypothesen H1 bis H3 bezüglich der Haupteffekte Klassenzugehörigkeit, Testzeitpunkt und Wechselwirkungen getestet.

Auf einen Überblick über die Mittelwerte (Tabelle 7.37) folgt eine Darstellung der Innersubjekteffekte (Tabelle 7.38), das sind die Testzeitpunkte, und der Zwischensubjekteffekte (Tabelle 7.39), das sind die Klassenzugehörigkeiten.

Deskriptive Statistiken

	Klasse	Mittelwert	SD	N
Spiegel-Prae	2	1,85	1,73	20
	8	3,10	1,91	10
	10	3,19	1,83	16
	Gesamt	2,59	1,88	46
Spiegel-Post	2	2,45	1,96	20
	8	4,50	2,27	10
	10	4,63	1,67	16
	Gesamt	3,65	2,17	46

Tabelle 7.37: Klassenweise Vergleiche der Prae- und Posttests; Thema Spiegel

Tests der Innersubjekteffekte					
Quelle	df	F	Sig.	Partielles Eta-Quadrat	Beobachtete Schärfe[a]
Testzeitpunkt	1	19,108	**0,000**	0,308	0,990
Testzeitpunkt * Klasse	2	1,317	**0,278**	0,058	0,269
Fehler (Testzeitpunkt)	43				
a. Unter Verwendung von Alpha = ,05 berechnet					

Tabelle 7.38: Effekte des Testzeitpunktes und Wechselwirkungen Testzeitpunkt-Klasse; Thema Spiegel

Tests der Zwischensubjekteffekte					
Quelle	df	F	Sig.	Partielles Eta-Quadrat	Beobachtete Schärfe[a]
Konstanter Term	1	164,832	0,000	0,793	1,000
Klasse	2	6,009	**0,005**	0,218	0,859
Fehler	43				
a. Unter Verwendung von Alpha = 0,05 berechnet					

Tabelle 7.39: Klassenmittelwerte werden verglichen und auf ihre Signifikanz geprüft; Thema Spiegel

Es ist zu erkennen, dass wiederum der Testzeitpunkt höchstsignifikant ist ($p < 0{,}001$), was für alle Klassen ein besseres Abschneiden im Posttest bedeutet. Die Wechselwirkung Testzeitpunkt und Klasse ist nicht signifikant, weist jedoch ein partielles $\eta^2 = 0{,}058$ auf, dessen Bedeutung aufgrund der kleinen Stichprobengröße hier noch zu prüfen ist.

Der klassenweise Unterschied (Tabelle 7.39) fällt auch hier hochsignifikant aus ($p = 0{,}05$), wieder mit einer moderaten Effektstärke (0,218).

Die in Abbildung 7.20 und die Abbildung 7.21 dargestellten Interaktionen lassen eine Interpretation beider Haupteffekte, Testzeitpunkt und Klasse, zu. Da in diesem Fall in beiden Diagrammen die Rangordnungen erhalten bleiben, sind beide Haupteffekte

dahingehend interpretierbar, dass sowohl der Testzeitpunkt als auch die Klassenzugehörigkeit für das Abschneiden im Wissenstest entscheidend sind.

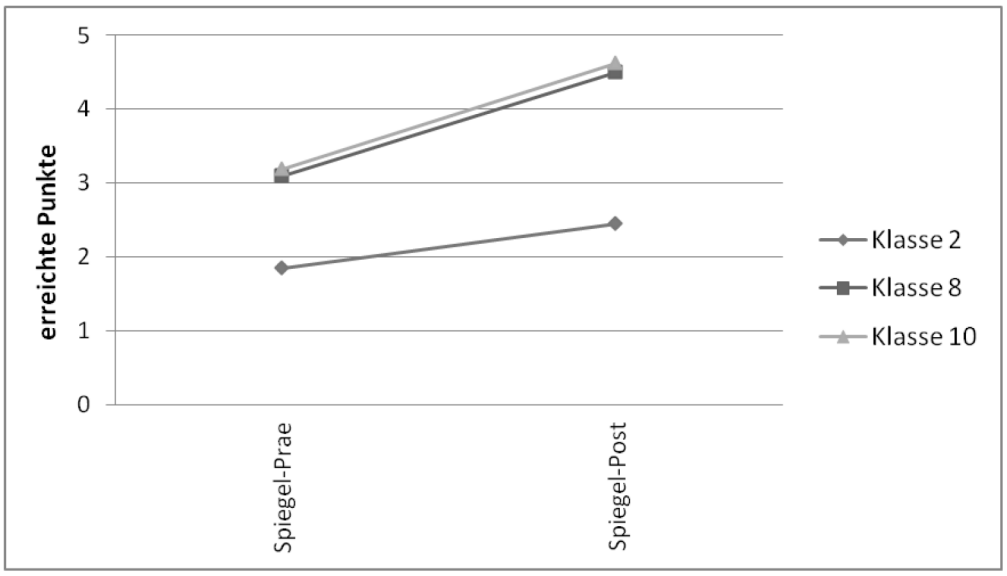


Abbildung 7.20: Darstellung des Haupteffektes Testzeitpunkt für die einzelnen Klassen; Thema Spiegel

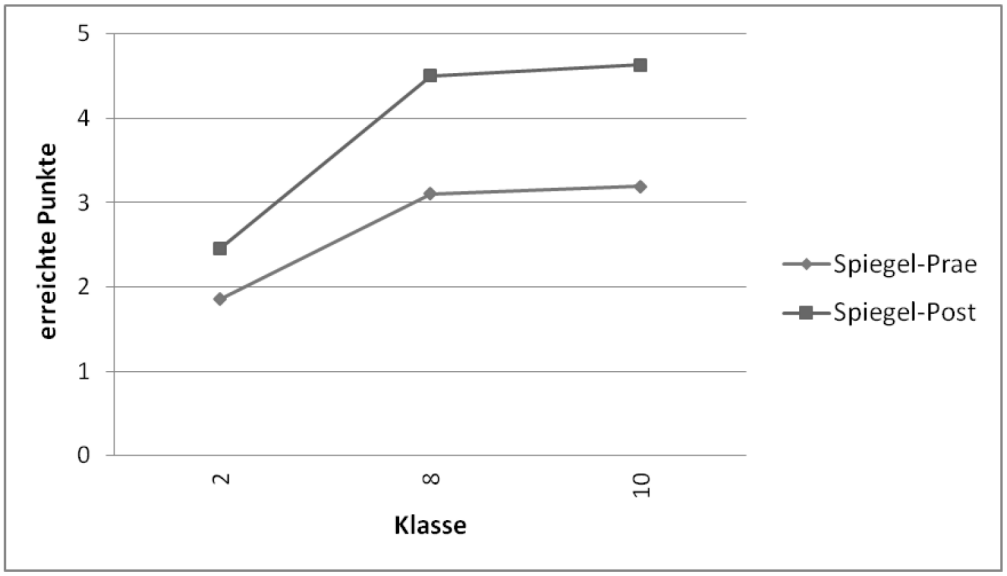


Abbildung 7.21: Darstellung des Haupteffektes Klasse; Thema: Spiegel

Mit anderen Worten: Die Posttests fallen hochsignifikant besser aus als die Praetests. Ein Unterschied im Praetest zwischen den Klassen bleibt auch im Posttest bestehen. Von einer Klasse, die a priori einen schlechteren Praetest als eine andere aufweist, ist zu erwarten, dass sie auch nach der Intervention durch CAPT einen schlechteren Posttest aufweisen wird.

7.3. Analysen zu den Follow-up Tests

In diesem Kapitel wird die ***Forschungsfrage 5:*** *„Wie nachhaltig ist die kognitive Verbesserung, die durch die CAPT-Intervention erzielt wurde?“* analysiert.

Methodik

Die Follow-up Tests wurden zwei bis fünf Wochen nach den Posttests durchgeführt. Trotz dieser Spannbreite kann man davon ausgehen, dass sich die Testbedingungen über diesen Zeitraum wenig verändert haben. Längere Abstände zwischen den Testzeitpunkten bedeuten allenfalls strengere Kriterien für die Persistenz des durch CAPT akquirierten Wissens (siehe Kapitel 4.1).

Aufgrund der Anzahl an fehlenden Werten und der Empfehlungen von Graham (2009) wurden fehlende Werte zu allen Testzeitpunkten für die hier präsentierten Vergleiche mit *Multipler Imputation* behandelt (vgl. Kapitel 4.5). Es wurden jeweils 20 Imputationen berechnet und anschließend über deren Ergebnisse gemittelt.

Prinzipiell würde sich für die Untersuchung dieser Fragestellung die einfaktorielle Varianzanalyse mit Messwiederholung anbieten. Eine der Voraussetzungen für deren Durchführung ist aber die Balanciertheit des Designs, d.h. es muss für jeden Testzeitpunkt und jede Person ein Messwert vorliegen. Sonst kann die Person nicht berücksichtigt werden. Das entspricht einer *casewise deletion*. Im Falle der vorliegenden Daten würde, diesem Kriterium entsprechend, eine große Anzahl an Fällen ausgeschlossen werden müssen. Daher wurde auf dieses Verfahren verzichtet. Stattdessen wurden jeweils zwei Testzeitpunkte mittels t-Test miteinander verglichen und die Ergebnisse dieser Vergleiche interpretiert.

Überblick über die Testergebnisse für Elektrizitätslehre und Optik

Abbildung 7.22 zeigt die mittleren Punktescores zu allen drei Testzeitpunkten (Praetest, Posttest, Follow-up Test) der jeweiligen Wissenstests zu den beiden Bereichen Elektrizitätslehre und Optik. Der Bereich Optik selbst, zerfällt wiederum in die Themen Schatten und Spiegel. Die Grafik soll als erste Orientierung für das Abschneiden in den Wissenstests dienen. Die absoluten Zahlen der drei mittleren Punktescores sind aufgrund der unterschiedlichen Itemanzahlen im jeweiligen Test nicht aussagekräftig und die Verbindungslinien dienen lediglich der besseren Übersicht. Es ist zu erkennen,

dass jedenfalls die Posttestscores über den Praetestscores liegen. Die Scores der Follow-up Tests liegen ebenfalls über denen der Praetests. Das dient als erster Hinweis auf die Wirkung von CAPT auch über die unmittelbare Intervention hinaus.

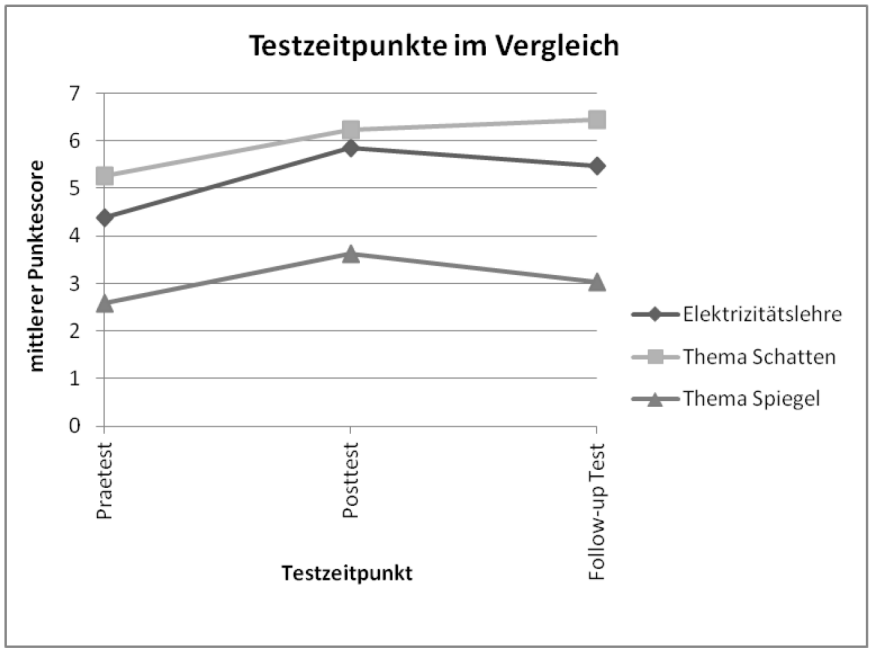


Abbildung 7.22: Mittlere Punktescores zu allen drei Testzeitpunkten (Praetest, Posttest und Follow-up Test) und den beiden Bereichen Elektrizitätslehre und Optik im Vergleich

Für die Elektrizitätslehre und das Thema Spiegel liegt der jeweilige mittlere Score des Follow-up Tests unter jenem des Posttests. Interessanter Weise liegt jedoch für das Thema Schatten der Follow-up Testscore über dem des Posttests und liefert somit den höchsten Wert.

Um eine mögliche Inferenz dieser Befunde auf die Population zu überprüfen, wurden die Mittelwerte von jeweils zwei Testzeitpunkten mittels t-Test verglichen (Tabelle 7.40).

Wie bereits in den vorangehenden Kapiteln berichtet, gibt es für alle drei getesteten Bereiche eine hochsignifikante Verbesserung vom Praetest zum Posttest ($p < 0{,}001$).

Posttest versus Follow-up Testergebnisse für Elektrizitätslehre und Optik

Um das „Vergessen“ der Inhalte der Intervention und der erarbeiteten Konzepte zu beurteilen wurden die Posttests mit den Follow-up Tests verglichen. Hier fallen die Ergebnisse der Tests zwar unterschiedlicher aus, aber in keinem der untersuchten Fälle ist ein hochsignifikanter Unterschied zwischen den beiden Testzeitpunkten zu finden.

Weder unterscheiden sich die geringeren Werte der Follow-up Tests hochsignifikant von denen der Posttests, noch ist für das Thema Schatten der höhere Wert im Follow-up Test signifikant höher als der Posttest. Allerdings sind in der Elektrizitätslehre die Follow-up Tests auf einem 5 %-Niveau signifikant und für das Thema Spiegel immerhin tendenziell signifikant.

Thema	Vergleich	Mittelwert	T	Signifikanz (2-seitig)
Elektrizitätslehre	Prae-Post	-1,45	-6,086	0,000
	Prae-Follow-up	-1,08	-4,488	0,000
	Post-Follow-up	0,37	1,982	0,048
Schatten	Prae-Post	-0,92	-4,271	0,000
	Prae-Follow-up	-1,175	-4,665	0,000
	Post-Follow-up	-0,225	-0,927	0,354
Spiegel	Prae-Post	-1,07	-4,205	0,000
	Prae-Follow-up	-0,462	-1,297	0,196
	Post-Follow-up	0,577	1,897	0,059

Tabelle 7.40: Mittlerer Wissenszuwachs (negativ) und mittlere Wissensabnahme (positiv) zwischen je zwei Testzeitpunkten

Was in Summe, bezogen auf den Testzeitpunkt des Follow-up Tests, von CAPT übrig bleibt, ist, dass für die Elektrizitätslehre eine hochsignifikante Verbesserung gegenüber dem Praetest auch nach einigen Wochen zu finden ist ($p < 0,001$). Die Effektstärke beträgt hierfür 0,36. Für das Thema Schatten fällt der Befund ähnlich aus, wobei hier sogar noch eine Steigerung gegenüber der Posttests zu finden ist. Die Effektstärke beträgt in diesem Fall 0,49. Eventuell kann die zusätzliche Steigerung im Follow-up Test als Hinweis auf Test-Retest Effekte zu werten sein.

Lediglich für das Thema Spiegel führt der Praetest – Follow-up Test Vergleich zwar zu einer bleibenden Zunahme im Wissen (0,46 Punkte), jedoch ist dieser Unterschied nicht signifikant ($p = 0,059$), wenn auch die kleine Übertretungswahrscheinlichkeit in Richtung einer Signifikanz deutet.

Vergleich nach Rollen für die Elektrizitätslehre

Aufgrund der teilweise hohen Anzahl an fehlenden Werten wurde für die Optik auf klassenweise Vergleiche verzichtet und für die Follow-up Tests werden somit nur die

oben berichteten Trends beschrieben. Da die Datenlage im Bereich der Elektrizitätslehre günstiger war, wurden hier, ähnlich wie bei den Prae-Post-Vergleichen, auch Analysen zu einem Zusammenhang zwischen Langzeitwirkung und Rolle angestellt.

Tabelle 7.41 zeigt Vergleiche zwischen jeweils zwei Testzeitpunkten, wobei das Sample nach Rollen aufgeteilt wurde.

Rolle	Vergleich	Mittelwert	T	Signifikanz (2-seitig)
Tutoren	Prae-Post	-1,29	-4,200	,000
	Prae-Follow-up	-0,74	-2,382	,017
	Post-Follow-up	0,55	2,090	,037
Tutees/Tutoren	Prae-Post	-2,33	-4,628	,000
	Prae-Follow-up	-2,08	-3,755	,000
	Post-Follow-up	0,25	0,671	,502
Tutees	Prae-Post	-0,83	-1,517	0,129
	Prae-Follow-up	-0,80	-1,688	0,091
	Post-Follow-up	0,03	0,071	0,944
aktive Rolle	Prae-Post	-1,61	-6,088	0,000
	Prae-Follow-up	-1,15	-4,149	0,000
	Post-Follow-up	0,46	2,124	0,034

Tabelle 7.41: Mittlerer Wissenszuwachs (negativ) und mittlere Wissensabnahme (positiv) in der Elektrizitätslehre zwischen je zwei Testzeitpunkten, sortiert nach Rollen

Die Vergleiche zwischen den Mittelwerten von Posttest und Follow-up Test zeigen, dass alle Schülergruppen im Mittel im Follow-up Test etwas weniger Punkte erreichen als im Posttest. Erstaunlicher Weise ist der Unterschied in der Gruppe der Tutees am geringsten, obwohl diese die kürzeste Intervention genossen haben. In keiner der Gruppen ist der Unterschied zwischen Posttest und Follow-up Test hochsignifikant. Für die Tutoren und die Schüler/innen in der aktiven Rolle ist er jedoch auf einem 5 % - Niveau signifikant. Zusammenfassend bedeutet das, vom Zeitpunkt des Posttests bis hin zum Follow-up Test wurde der Wissensstand der Schüler/innen, unabhängig von der Rolle, die sie im Tutoring Prozess inne hatten, etwas geringer. In allen Fällen bleibt der Punktescore über dem des Praetests, was für die Persistenz des Wissens spricht.

Vergleicht man die Praetests mit den Follow-up Tests lässt sich eine Aussage darüber machen, wie anhaltend die Wissensvermittlung durch die Intervention im Vergleich zum Ausgangszustand ist. In Tabelle 7.40 wurde bereits gezeigt, dass für alle getesteten Personen in der Elektrizitätslehre, unabhängig von ihrer Rolle, eine hochsignifikante Verbesserung erhalten bleibt. Den wesentlichen Beitrag dazu liefern die Schüler/innen in der aktiven Rolle. Innerhalb dieser Gruppe wiederum tragen vor allem die Schüler/innen in der Doppelrolle (Tutees/Tutoren) dazu bei. Für die Tutoren bleibt immer noch eine Steigerung in ihrem Wissen übrig, die etwa 60 % über dem Niveau der Praetests liegt, sie ist jedoch gerade nicht mehr signifikant. In diesem Fall kann es aber auch sein, dass die fehlende Signifikanz auf die reduzierte Samplegröße zurückzuführen ist.

7.4. Nicht-kognitive Parameter und Testergebnisse

In diesem Abschnitt soll Forschungsfrage 7 behandelt werden: Welche Zusammenhänge ergeben sich zwischen nicht-kognitiven Parametern und Testergebnissen?

Die motivationalen Parameter, die getestet wurden, wurden mit folgenden Fragebogeninstrumenten erhoben (vgl. Kapitel 5.3):

- intrinsische Motivation mit dem Fragebogen zum Lernen im Fach (F. H. Müller, et al., 2007)
- die Leistungsmotivation mit dem Fragebogen zur Erfassung der aktuellen Motivation (Rheinberg, et al., 2001), FAM
- die allgemeine Selbstwirksamkeitserwartung mit dem Fragebogen zur SWE (Schwarzer, 1993)
- Aspekte zur intrinsischen Motivation mit dem selbst entwickelten Motivationsfragebogen auf Basis des IMI (Korner, et al., 2012) – Fassung mit den besten 23 Items

Der Fragebogen zur SWE wurde lediglich im zweiten Projektjahr zu den Themen der Optik eingesetzt, da es Hinweise gab, dass das selbst entwickelte Fragebogeninstrument vielleicht nicht reliabel messen würde.

Auf Basis der im Vorfeld bereits diskutierten Literatur, gibt es fundierte Belege, dass jeder der oben genannten Punkte Einfluss auf die Lernleistung und somit auf die Testergebnisse hat. Daher wurden Korrelationen und Signifikanzen zur Bewertung der Stärke dieser zu erwartenden Zusammenhänge berechnet. Die Berechnung der Korrelationen wurde, gemeinsam mit der Theorie im Hintergrund, als ausreichend erachtet.

Im Falle der Elektrizitätslehre wurde im Vorfeld bereits gezeigt, dass die Praetests sehr homogen in den getesteten Parametern waren. Daher wurden ausschließlich Pearson Korrelationen der motivationalen Parameter zum Posttest-Ergebnis berechnet. Konkret wurden für den FAM und das selbst entwickelte Fragebogeninstrument die Korrelationen zwischen Posttest und Summe der jeweiligen Itemantworten berechnet. Im Falle des *Fragebogens zum Lernern im Fach* wurde der SDI, der *Self Determination Index*, wie von den Autoren vorgeschlagen, herangezogen.

Korrelation Signifikanz	Posttest	Lernen im Fach	Motivation	SDI
Posttest	-	0,051 0,560	**0,350** **0,001**	-0,004 0,966
Lernen im Fach	0,051 0,560	-	0,210 0,051	-0,170 0,042
Motivation	**0,350** **0,001**	0,210 0,051	-	-0,076 0,462
SDI	-0,004 0,966	-0,170 0,042	-0,076 0,462	-

Tabelle 7.42: Korrelationen und Signifikanzen für die Elektrizitätslehre

Tabelle 7.42 zeigt die berechneten Korrelationen und deren Signifikanzen. Die einzige interessante Korrelation eines dieser nicht-kognitiven Parameter zum Posttest-Ergebnis ist jene mit dem selbst entwickelten Fragebogeninstrument, mit r = 0,350 und auf einem Signifikanzniveau von p = 0,001. Nun ist die Korrelation zwar hochsignifikant, jedoch ist der Zusammenhang mit einem r von 0,350 nicht besonders stark. Dennoch kann man folgern, dass motivierte Schüler/innen im Posttest etwas besser abschneiden werden. Alle anderen Korrelationen zum Posttest sind schwach und bei weitem nicht signifikant.

Interessant ist die Tatsache, dass das selbst entwickelte Motivationsinstrument mit $r = 0{,}210$ und $p = 0{,}051$ mit dem *Fragebogen zum Lernen* im Fach korreliert. Dieser Zusammenhang ist zwar nicht stark, dennoch kann er als externer Hinweis auf die Validität des Motivationsinstrumentes aufgefasst werden, da beide Instrumente auf der *Self Determination Theory* basieren.

Im Falle der Themen Spiegel und Schatten aus der Optik wurden die Zuwächse (*gains*) zur Berechnung der Korrelationen zu nicht-kognitiven Parametern herangezogen, da die Ausgangslage in den Praetests sehr unterschiedlich war. Die nicht-kognitiven Parameter waren Lernen im Fach, Selbstwirksamkeitserwartung, Motivation und der SDI. Für diese Variablen ergab sich keine einzige Korrelation, die größer war als 0,18 und das auf einem weit von jeder Signifikanz entfernten Niveau.

Ein Zusammenhang zwischen dem Geschlecht und dem Abschneiden im Posttest wurde für die Elektrizitätslehre geprüft. Für die Optik wurden derartige Analysen nicht durchgeführt, da hier bereits die Praetests sehr unterschiedlich waren. Interessanter Weise gibt es in der Elektrizitätslehre zwischen Posttest und Geschlecht nur eine sehr schwache Korrelation ($r = -0{,}073$), die nicht signifikant ist ($p = 0{,}383$). In anderen Worten: Das Posttestergebnis ist im Sample nicht vom Geschlecht der Schüler/innen anhängig, insbesondere schneiden Mädchen bei CAPT nicht schlechter ab als Jungen. Dieses Ergebnis Elektrizitätslehre ist somit auch in der Population zu erwarten.

8. Diskussion

Die hier vorgestellte Arbeit versuchte verschiedene Aspekte des sehr komplexen Cross-Age Peer Tutorings abzubilden. Hauptaugenmerk lag dabei auf dem Wissenserwerb der Schüler/innen in unterschiedlichen Bereichen, aber auch auf ihrer Motivation. Dem entsprechend gab es eine lange Phase der Datenerhebung, die zwei Schuljahre in Anspruch nahm.

In der nun folgenden Diskussion werden, den Forschungsfragen entsprechend, drei Hauptziele diskutiert. Im ersten Abschnitt soll darauf eingegangen werden, ob sich die Methode CAPT überhaupt für den Physik eignet und welche Schülergruppen, wenn überhaupt, davon profitieren. Die Erhebungen dazu geschahen inhaltlich anhand einiger Themen aus der Elektrizitätslehre. Der Hauptgrund für diese Wahl lag darin, dass die Elektrizitätslehre ein zentrales Gebiet der Physik ist, sowohl für die Fachwissenschaft, als auch für die hier untersuchte Schülerpopulation. Darüber hinaus gibt es hier gut erforschte Schülervorstellungen und es lag für die Untersuchung ein valides Testinstrument zur Elektrizitätslehre (Urban-Woldron & Hopf, 2012) vor.

Im zweiten Abschnitt wird diskutiert wie sich CAPT auch auf andere Themengebiete anwenden lässt. Dazu wurden Inhalte aus der Optik gewählt, die einerseits auch ein wesentlicher Teil des Lehrplanes der Sekundarstufe 1 sind und zu der es andererseits ebenfalls gut erforschte Schülervorstellungen gibt. Im dritten Abschnitt soll das Unternehmen, ein Instrument zur Erfassung differenzierter Motivationsformen anzupassen und zu verwenden kritisch beleuchtet werden.

Zunächst soll noch einmal betont werden, dass CAPT sich als konstruktivistischer Unterrichtsansatz versteht (vgl. Kapitel 2.1.3). Mentoring und Tutoring stellen die hierfür benötigten Lernumgebungen dar, die eine Eigenkonstruktion von Konzepten ermöglichen sollen, sowohl für die Tutoren als auch für die Tutees. Im Zentrum des Geschehens stehen die Lernenden. Besonders hervorzuheben ist neben der Fokussierung auf Schülervorstellungen die Möglichkeit der sozialen Interaktionen, die das Mentoring und das Tutoring bieten. Hier passiert Wissenskonstruktion von Beginn an nicht isoliert, sondern im sozialen Kontext. Beim Tutoring ist das augenscheinlich klar, da in dem gewählten eins-zu-eins Setting intensive Kommunikation zwischen den

Tutoren und den Tutees stattfindet. Aber auch das Mentoring unterstützt Lernen im sozialen Kontext, da die Schüler/innen hier in der Kleingruppe arbeiten. Viel mehr aber wird der soziale Kontext durch die Tatsache angesprochen, dass den zukünftigen Tutoren von Beginn an vermittelt wird, dass sie in absehbarer Zeit ihr Wissen an ihre Tutees weitergeben sollen. Somit erleben sie die Herausforderung parallel zum eigenen Wissenserwerb sofort über einen möglichen Transfer nachzudenken.

Die für konstruktivistische Lernumgebungen zentrale Berücksichtigung des Vorwissens und der Schülervorstellungen war auch Kernpunkt im Aufbau des Mentoring, das um bereits aus der Literatur bekannte Vorstellungen konstruiert wurde. Passend zu den Schülervorstellungen gab es jeweils Aufgaben, die auf Basis der P-O-E-Strategie zu lösen und zu diskutieren waren. Dabei war das Arbeitstempo und somit der Lernfortschritt individuell wählbar, individuelles und selbstgesteuertes Lernen wurden somit unterstützt. Innerhalb der Diskussionen fanden auch Aushandlungsprozesse zu möglichen Erklärungsmodellen statt, was in einer parallel zu dieser Arbeit durchgeführten Studie gezeigt werden konnte (Trinkl, 2012). Diese Aushandlungsprozesse sind in der konstruktivistischen Auffassung vom Lernen der Vorgang, bei dem Wissen *konstruiert* wird.

Auch zwischen Tutoren und Tutees führte die CAPT Intervention zu Aushandlungsprozessen über mögliche Erklärungskonzepte zu den Aufgaben (Trinkl, 2012). Eng damit verbunden ist nicht nur die Diskussion über die Konzepte des Anderen, sondern die Reflexion über die eigenen Konzepte. Dieser Prozess unterstützt neben der eigentlichen Konzeptentwicklung auch die Entwicklung metakognitiver Fähigkeiten. Diese wiederum sind laut der konstruktivistischen Sichtweise notwendig, um besser und schneller im Sinne von Aufschnaiter (1992) neue Erfahrungsbereiche konstruieren zu können.

Eigenständigkeit und Selbstverantwortung wurden bestmöglich unterstützt, indem die zukünftigen Tutoren einerseits inhaltlich eine gewisse Wahlfreiheit hatten, andererseits als Expert/innen für das Lernen ihrer Tutees angesprochen wurden und somit Verantwortung für deren Lernprozesse übernehmen sollten. Damit verbunden ist das Einnehmen multipler Perspektiven: Inhalte, nehmen wir als Beispiel die Größe von Schlagschatten, wurden aus der persönlichen Sicht des Lernenden betrachtet und

danach aus der Sicht des Vermittelnden, mit Blick auf potenzielle Lernschwierigkeiten für Tutees. Die Sicht des Vermittelenden selbst beinhaltet zusätzlich, dass unterschiedliche Zugänge und Kontexte zum Schlagschatten gefunden und erprobt werden.

Beiden, Tutees und Tutoren, wurden nach Möglichkeit bedeutungsvolle Inhalte angeboten, also solche, die sich an ihrer Lebenswelt orientierten oder die sie aus ihrer Lebenswelt kannten. So wurden in der Optik z.B. Schatten- und Spiegelphänomene besprochen, die jeder aus dem Alltag kennt. Diese wurden anhand von Experimenten untersucht, wobei die Materialien zum Großteil Alltagsgegenstände wie gewöhnliche Taschenlampen, Filzstifte, Legofiguren, Spielzeugautos oder Tiere aus Überraschungseiern waren. Auch im Bereich der Elektrizitätslehre wurde auf gewöhnliche Batterien, Kabel und Lämpchen zurückgegriffen und keine komplizierten, Blackboxes ähnlichen Bausteinsysteme verwendet, um einen größtmöglichen Alltagsbezug herzustellen.

Die Ergebnisse aus der Elektrizitätslehre, wie auch der Optik legen nahe, dass CAPT in der Lage ist, die oben angeführten Aspekte konstruktivistisch gestalteter Lernumgebungen zu unterstützen und daher zu zufrieden stellenden Ergebnissen zu führen. Zur Beantwortung der Forschungsfrage 1 (siehe S. 45) konnte die Elektrizitätslehre betreffend gezeigt werden, dass die mittlere Effektstärke der Intervention über alle Schüler/innen bei 0,46 liegt, was Hattie (2009) als mittelstarken Effekt bezeichnet, der bereits in der *„zone of desired effects“* liegt. Die Optik betreffend lagen die mittleren Effektstärken für das Thema Schatten bei vergleichbaren 0,49, für das Thema Spiegel sogar bei 0,62, was Hattie bereits als großen Effekt charakterisiert. Das kann in Relation zu Effekten von Lehrer/innen[31] gesetzt werden, die Hattie mit 0,20 bis maximal 0,40 angibt.

Das ist umso erstaunlicher, da bei den Spiegeltests die Stichprobengröße mit N_{Sp} = 46 Probanden gering war. Trotzdem ist hier ein höchstsignifikanter Effekt mit einer interessanten Effektstärke zu beobachten. Hätte man eine a priori Stichprobenplanung

[31] Diese Angabe bedeutet, dass Lehrer/innen typischer Weise einen Lernfortschritt bei Schüler/innen von 0,20 bis 0,40 Standardabweichungen bewirken können, wenn man sie vor eine Klasse stellt. Effekte dieser Größe werden von Hattie als durchschnittlich bezeichnet (Hattie, 2009, S. 17).

durchgeführt, wären für eine ausreichende statistische Untermauerung solcher Effekte Stichproben mit etwa 60 Probanden vonnöten[32]. Eine a priori Stichprobenplanung war nicht möglich, da die Größenordnung der zu beobachtenden Effekte dafür schon im Vorhinein bekannt sein hätte müssen. Dass sich trotz der geringen Stichprobengröße ein derart starker Effekt zeigt, spricht für die Wirksamkeit der Unterrichtsmethode CAPT.

CAPT hat somit eine Lernwirksamkeit, die deutlich über der von gewöhnlichem Unterricht liegt. Das ist insofern zufriedenstellend, da die Interventionen je nach Rolle der Schüler/innen im Tutoringprozess zwischen etwa 40 Minuten und 130 Minuten dauerten und somit eher als kurz zu bezeichnen sind.

Die Analysen zu Forschungsfrage 1 legen zwei Dinge offen: Zum einen zeigen sie, dass Schüler/innen der Sekundarstufe 1 Lernerfolge aufgrund der CAPT-Intervention aufweisen. Das war aufgrund vorangegangener Studien zu CAPT, die sich eher auf Grundschüler/innen oder auf Ältere, Student/innen an Colleges, beziehen, nicht a priori klar. Zum anderen konnte gezeigt werden, dass CAPT auch in Physik zu verbesserten Testergebnissen in dieser Altersgruppe führt. Dies ergänzt die Literaturangaben, da die Verbesserung der Testergebnisse zwischen Prae- und Posttest dahin gehend interpretiert werden kann, dass CAPT zu belastbareren – wissenschaftlich korrekten – Konzepten führt und somit einen Konzeptwechsel unterstützt, wenn nicht sogar initiiert. Auf Basis der vorliegenden Untersuchungen sind jedoch keine Aussagen darüber möglich, auf welchen Mechanismen dieser Konzeptwechsel basiert. Eine konkrete Beforschung der Lernprozesse mit qualitativen Methoden wäre hier sicher in Zukunft interessant und hilfreich.

Bei der Datenauswertung zur Elektrizitätslehre wurde bewusst drauf verzichtet nach Konzepten auszuwerten, obwohl der verwendete Test versprach dies leisten zu können. Die Ursache dafür war die geringe Zahl der zu testenden Konzepte und die Tatsache, dass für die Optik kein derartiges Testinstrument zur Verfügung gestanden hatte, was eine mögliche Vergleichbarkeit der Daten beider Studienteile beeinträchtigt hätte. Wenn man auf diese Vergleichbarkeit jedoch einen Augenblick lang verzichten möchte und doch die konzeptuelle Zählart des Test zum Verständnis der Elektrizitätslehre (Urban-

[32] Diese Abschätzung erfolgte aufgrund einer post-hoc Analysemittels G*Power auf Basis einer geschätzten Effektstärke von 0,6 und einer Irrtumswahrscheinlichkeit von $\alpha = 0,05$.

Woldron & Hopf, 2012) anwendete, können die Ergebnisse aus der summativen Zählart bestätigt werden. Signifikante Unterschiede bleiben bestehen, wenn auch auf einem etwas reduzierten Signifikanzniveau. Das spricht noch einmal dafür, dass mit der Methode CAPT ein Konzeptwechsel in die Wege geleitet werden kann. Zum anderen zeigt es auch, dass die Effekte auf das Wissen, das durch CAPT vermittelt wird, praktisch so bedeutsam sind, dass sie keines elaborierten Tests bedürfen, um nachgewiesen werden zu können und auch nicht von einer speziellen Art der Auswertung abhängen.

Für den Schulalltag lassen sich die Ergebnisse so interpretieren, dass Schüler/innen, egal in welcher Rolle, von CAPT profitieren. Auch bedarf es zunächst keiner speziellen diagnostischen Maßnahmen, um passende Paarungen zu finden. Das erleichtert eine schnelle und kurzfristig organisierbare Durchführung mit ganzen Klassen.

Was den Einfluss des Mentorings auf die Posttest-Ergebnisse betrifft, so wurde dieser nicht explizit im Rahmen der Untersuchungen bestimmt. Jedenfalls ist aufgrund der großen Effektstärken von CAPT davon auszugehen, dass die Methode als solche wirkt und nicht bloß das Mentoring, das zwar unterschiedliche Ziele hat, aber zumindest teilweise mit direkter Instruktion verglichen werden kann. Der Einfluss des Mentoring wurde aus dem Grund nicht genauer untersucht, da es zunächst darum ging festzustellen, ob CAPT im physikalischen Kontext überhaupt wirkt. Diesen Einfluss zu quantifizieren wäre sicher eine lohnenswerte Aufgabe für eine Folgestudie und es ließe sich mit einem relativ geringen Aufwand in einem Vergleichsgruppendesign realisieren. Dass das Mentoring aber einen positiven Einfluss auf die Lernergebnisse hat, sollte sich schon alleine aus der Tatsache ergeben, dass es an Schülervorstellungen orientiert war. Unterricht, der sich an Schülervorstellungen orientiert, ist immer effektiver als einer, der darauf keine Rücksicht nimmt (Duit & Treagust, 2003).

Die Beantwortung der Forschungsfragen 2, 3 und 4 nach der detaillierten Wirksamkeit von CAPT konnte korrekt nur für die Elektrizitätslehre durchgeführt werden. Der Hauptgrund lag darin, dass in der Elektrizitätslehre die Praetests bezüglich der getesteten Parameter homogen waren, in der Optik jedoch nicht. Auf diese vorbereitenden Analysen wurde sehr viel Augenmerk gelegt, damit die daraus folgenden Inferenzen fundiert sind. In der Elektrizitätslehre zeigte sich, dass sich die Klassen im Vorwissen nicht signifikant unterschieden, egal, ob es die AHS-Klasse (Klasse 6) war, die

Hauptschulklassen, oder die Klasse (Klasse 1a), die bereits im Vorjahr Unterricht zu diesem Thema hatte. Auch konnte kein Einfluss der unterschiedlichen Altersstufen festgestellt werden. Ältere Schüler/innen aus der 8. Schulstufe zeigten kein signifikant unterschiedliches Vorwissen im Vergleich zu jenen aus der 6. Schulstufe. Ebenso erfolgte die Zuteilung der Rollen, Tutoren-Tutees-Doppelrolle, nicht auf Basis des Vorwissens. Eine derartige ANOVA war nicht signifikant, woraus geschlossen werden kann, dass die Rollenzuteilung tatsächlich zufällig und nicht nach Vorwissen erfolgte.

Im Bereich der Optik arbeiteten die Klassen zu zwei unterschiedlichen Themen, zum Schatten und zum Spiegel mit jeweils unterschiedlichen Testinstrumenten. Innerhalb dieser Klassen wurde dann nach den drei Rollen aufgeteilt. Daraus ergab sich für jedes einzelne Thema und für jede Rolle eine geringe Probandenzahl, die signifikante Aussagen für die Population nicht mehr möglich machte. Darüber hinaus unterschieden sich die Klassen signifikant im Vorwissen, was zuverlässige Inferenzen nicht mehr möglich machte. Zunächst fiel die AHS-Klasse (Klasse 6) aus dem Rahmen, da sie nicht nur eine viel längere Interventionsdauer hatte, sondern auch beide Themen bearbeiten konnte. Das bedeutete, dass diese Klasse eine Vielzahl an Lerngelegenheiten hatte und unter mehreren Perspektiven die Problemstellungen bearbeiten musste, was in diesem Fall offenbar sehr wohl zu Effekten führte. Auch handelte es sich hier um die einzige Klasse aus der AHS, alle andern Klassen waren Hauptschulklassen. Aufgrund der Summe dieser Faktoren wurde Klasse 6 in den weiteren Analysen zur Optik nicht berücksichtigt. Mehr oder minder großen Inhomogenitäten zwischen den verbleibenden Klassen wurde insofern Rechnung getragen, dass Analysen auf Klassenebene durchgeführt wurden und nicht mehr auf Basis der unterschiedlichen Rollen in Tutoringprozess, wie dies in der Elektrizitätslehre der Fall war.

Forschungsfrage 2 beschäftigte sich mit der Wirkung von CAPT auf die *Tutoren*. Nach den Ergebnissen vorangegangener Studien (Fogarty & Wang, 1982; Robinson, et al., 2005; Topping, 1996) erschien es sinnvoll die Ergebnisse der Tutor/innen genau unter die Lupe zu nehmen. Untersucht wurde der Effekt auf die Gruppe jener Schüler/innen, die ausschließlich als Tutoren wirkten, nicht aber auf die Gruppe derer, die in der Doppelrolle Tutees/Tutoren waren. Dies geschah, um das Ergebnis durch mögliche Dosiseffekte aufgrund unterschiedlich langer Interventionsdauern nicht zu verzerren.

Mit einer Effektstärke von 0,42 ist der Effekt zwar etwas kleiner als der mittlere Effekt über das gesamte Sample, aber dennoch zufriedenstellend. Die Ursache für die etwas geringere Effektstärke mag durchaus in der geringeren Samplegröße (N = 91) liegen, da damit die Standardabweichung wächst. Die erzielte Effektstärke bestätigt nicht nur die Resultate vorangegangener Studien, sondern erweitert deren Ergebnisse auf die Gruppe der Schüler/innen aus der Sekundarstufe und für Inhalte aus der Physik.

Dieser Effekt auf die Tutoren lässt einen veränderten Blickwinkel auf Peer Tutoring Prozesse hinsichtlich einer Implementierung in reguläre Klassen zu. Ein oft berichtetes Hemmnis für die Einrichtung derartiger Arbeitsformen in Schulen und in den Schulunterricht ist, dass kaum jemand daran denkt, dass eben die Tutoren auch vom Erklären profitieren. Die an der Studie beteiligten Klassenlehrer/innen bildeten da keine Ausnahme und glaubten, dass lediglich oder vor allem die Tutees profitierten. Das würde wiederum die Frage aufwerfen, warum sich Tutoren dem Prozess unterwerfen sollen, wenn sie keine Vorteile daraus zögen und sogar noch Zeit investieren müssten. Die Lehrer sähen sich gefordert, für die Tutoren ein Belohnungssystem zu entwickeln. Das kann aufgrund der oben besprochenen Ergebnisse als hinfällig betrachtet werden. Es kann entgegen gehalten werden, dass die Tutoren von einem Tutoring Prozess jedenfalls profitieren, auch wenn die getesteten Konzepte vermeintlich sehr einfach sind. Ein Belohnungssystem erübrigt sich somit und der Zeitaufwand für die Tutoren lohnt sich jedenfalls. Für den Schulalltag bedeutet das, dass Cross-Age Peer Tutoring Sequenzen jedenfalls ohne begleitende Maßnahmen übernommen werden können und Lehrer/innen davon ausgehen können, dass auch die Tutoren ohne spezielles Belohnungssystem davon profitieren.

Nachdem bisher gezeigt werden konnte, dass CAPT für alle Schüler/innen funktioniert und dass im Speziellen die Tutoren davon ebenfalls profitieren, versuchte Forschungsfrage 3 zu ergründen, ob und wenn ja, welche Unterschiede es im Wissen der Schüler/innen nach der CAPT Intervention gab und welche Ursachen dafür festgemacht werden können. Dazu wurde das Sample nach unterschiedlichen Gesichtspunkten wie der Klassenzugehörigkeit, der Rollenverteilung und den Altersstufen unterteilt. Ein F-Test fiel jedoch nur für die Unterteilung nach Rollen höchstsignifikant aus. An dieser Stelle ist es notwendig sich in Erinnerung zu rufen, dass die Praetests hinsichtlich der

Rollenzuteilung, wie auch der Altersstufen und der Klassenzugehörigkeiten, homogen waren. Das lässt die Folgerung zu, dass es tatsächlich an den unterschiedlichen Rollen während der CAPT-Intervention lag, dass die Schüler/innen im Posttest unterschiedliche Ergebnisse zeigten und nicht an irgendwelchen anderen Effekten.

Da es bereits starke Hinweise aus der Literatur darauf gab, dass die aktive Rolle als Tutor/in zu stärkeren Effekten führt als die passive Rolle der Tutees, wurden geplante Kontrasttests durchgeführt um die Effekte des Treatments, also den systematischen Teil der Varianz, näher zu untersuchen. Dazu wurden alle möglichen Kombinationen von Gruppen (Tutoren, Tutoren/Tutees, Tutees) miteinander verglichen. Lediglich der Vergleich der Schüler/innen in der aktiven Rolle, die Tutoren und Tutees/Tutoren inkludiert, mit denen der passiven Rolle (Tutees) war hierbei signifikant. Das lässt die Folgerung zu, dass die aktive Rolle die entscheidende ist, was die Wirksamkeit von CAPT betrifft. Daher ist es nicht nur so, dass *auch* die Tutoren von dieser Methode profitieren, sondern *vor allem* diese.

Kritiker können an dieser Stelle natürlich einwenden, dass die Unterschiede nicht durch Prozesse des Erklärens im Zuge von CAPT hervorgerufen wurden, sondern allein mit dem Mentoring zu erklären sind. Dem ist entgegen zu halten, wie auch schon bei den Analysen zu Forschungsfrage 1, dass die gezeigten Effekte mit einer Effektstärke von 0,62 auftreten. Es ist äußerst unwahrscheinlich (Duit & Treagust, 2003; Hattie, 2009), dass die relativ kurze Intervention des Mentoring derartige Effekte hervorruft. Dennoch hat das Mentoring sicherlich auch einen Einfluss auf die Größe des Effekts und es ist einmal mehr ein Grund, das Mentoring in einer Folgestudie genauer unter die Lupe zu nehmen. Eine mögliche interessante Forschungsfrage wäre es herauszufinden, ob das Mentoring auch außerhalb der Schulzeit Diskussionsprozesse in Gang bringt. Derartige Prozesse wären Teil eines konstruktivistisch selbstgesteuerten, individuellen Lernens und somit dem Potenzial der Unterrichtsmethode CAPT zuzurechnen. Weitere Dinge, die zwischen Mentoring und Tutoring geschehen hätten können und die nicht der Leistungsfähigkeit der Methode CAPT zuzuschreiben gewesen wären, wurden nach Möglichkeit ausgeschlossen. So war mit den beteiligten Lehrer/innen ausgemacht, zwischenzeitlich nichts zu den Themen der Intervention mit den Schüler/innen zu besprechen, also keine „Nachhilfe“ zu geben.

Darüber hinaus ist allerdings einzuwenden, dass Test-Retest Effekte eine mögliche Quelle für verzerrte Ergebnisse darstellen. Solche Effekte wurden im Rahmen der hier beschriebenen Untersuchungen nicht erforscht. In Folgestudien, die nicht nur die prinzipielle Wirksamkeit von CAPT feststellen wollen, sondern genauere Analysen zur Wirksamkeit anstreben wäre es sicherlich von Vorteil auch diese zu berücksichtigen.

Die letzte Untersuchung innerhalb der Forschungsfrage 3 beschäftigt sich mit möglichen Dosiseffekten, die aufgrund der unterschiedlichen Interventionsdauer zwischen Tutoren und Tutee/Tutoren auftreten könnten. Schüler/innen in der Doppelrolle hatten ein Tutoring mehr als jene, die nur als Tutoren agierten. Hier ist der durchgeführte Kontrasttest nicht signifikant. Das bedeutet, dass die Effekte des Tutorings nicht auf die Dauer zurückgeführt werden können und insbesondere die Unterschiede zwischen aktiver und passiver Rolle ebenfalls nicht der Interventionsdauer zuzuschreiben sind. Dieses Ergebnis mag insofern nicht überraschen, da sich jeder Teil der Intervention (Mentoring, Tutoring) an Konzepten orientierte und dieselben Konzepte angesprochen wurden. Eventuell hätte es einen Effekt der Dauer gegeben, wenn eine größere Zahl an unterschiedlichen Konzepten Gegenstand der Interventionen gewesen wäre. In unserem Fall wurde nur eine geringe Zahl an Basiskonzepten adressiert, was zu keinen Dosiseffekten führte.

Hinweise darauf, dass die Rolle der entscheidendere Faktor sein kann, liefert auch Hattie (2009) in seiner Studie, wenn er erklärt, dass im Hinblick auf der Ergebnis die Qualität der Instruktion essentiell ist, der Einfluss der *time-on-task* hingegen weniger als durchschnittlich. Cohen et al. (1982) und Robinson et al. (2005) berichten, dass kürzere Tutoring-Programme oft effektiver sind als längere, falls sie hochstrukturiert ablaufen. Für die hier vorgestellte CAPT-Intervention gehen wir davon aus, dass sie hochstrukturiert abgelaufen ist und aus diesem Grund wirksam ist, während die *time-on-task* einen untergeordneten Einfluss hat.

Diese Einschätzungen konnten auch mit Hilfe multipler linearer Regressionen bestätigt werden, da sich darin die Interventionszeit, abhängig von der Codierung, meist als nicht signifikanter Parameter herausstellte. Sie wird erst dann signifikant, wenn man wirklich die Minuten der *time-on-task* kodiert, alle Mentorings und Tutorings zusammen. Jedoch stellt sich dabei die Frage nach der Sinnhaftigkeit einer derartigen Rechnung, da eine

Kodierung in Minuten der unterschiedlichen Qualität der Interventionsteile nicht gerecht wird.

Im Zuge der Forschungsfrage 4 wurde versucht die obigen Ergebnisse einerseits zusammenzufassen, andererseits die Bedeutung einzelner Parameter als Prädiktoren für die Posttest-Ergebnisse des CAPT quantitativ in Relation zu setzen. Das ermöglichte eine Schätzung der Posttest-Ergebnisse für die Population. Dazu wurden verschiedene Modelle auf Basis multipler linearer Regressionen (MLR) gerechnet. Für die Berechnung der abhängigen Variable *Posttest* wurden unterschiedliche Prädiktoren hinzugenommen, für die jedoch aus der Literatur, den vorangegangenen Analysen und entsprechend ihrer Korrelationen mit der Variable *Posttest* ein Einfluss plausibel erschien. Es handelte sich daher um ein theoriegeleitetes Vorgehen und nicht um eine rein explorative Analyse. Die Methode für die Hinzunahme von Prädiktoren erfolgte in Blöcken nach der Einschlussmethode (*forced entry*).

Modell 1 und 2 zeigen, worauf schon die AVOVAS und t-Tests zu den Forschungsfragen 1 bis 3 hingewiesen haben: Neben dem Vorwissen, also dem Praetest-Ergebnis, ist die aktive Rolle entscheidend für ein gutes Abschneiden im Posttest. Die Variable *Ist Tutor* ist hochsignifikant ($p = 0{,}003$) und das zugehörige standardisierte Beta beträgt mehr als 0,2. Mit der Hinzunahme dieser Variable steigt auch der Determinationskoeffizient (korrigiertes R^2) von 0,175 auf 0,222 (vgl. **Tabelle 7.21**). Das ist ein recht beachtlicher Anteil an aufgeklärter Varianz für ein Modell, das lediglich zwei Prädiktoren enthält.

Vergleicht man die vier präsentierten Modelle (Tabelle 7.22) so fällt auf, dass das Modell 3 die kleinste Konstante (1,428) besitzt. Diese Konstante (*intercept*) gibt den erwarteten Posttestscore an, ohne den Beitrag irgendeines Prädiktors zu berücksichtigen. Daher ist das Modell mit der geringsten Konstante zu präferieren, da der unsystematische Anteil, der durch die Prädiktoren nicht geschätzt werden kann, hier minimal ist. Alle weiteren Punktezuwächse können mit Hilfe der Prädiktoren (Ergebnis beim) *Praestest*, der aktiven Rolle (*Ist Tutor*) und der *Muttersprache* geschätzt werden. Alle Prädiktoren sind für Modell 3 signifikant mit einer Übertretungswahrscheinlichkeit von $p < 0.009$ und zufrieden stellenden standardisierten Betas. Das kann gemeinsam mit der Tatsache, dass die Prädiktoren theoriegeleitet in das Modell aufgenommen wurden, dahingehend interpretiert werden, dass sie tatsächlich auch praktische Bedeutung haben. Dies

spiegelt sich auch in einer aufgeklärten Varianz (korrigiertes R^2) von 0,254 wider, die als gut bewertet werden kann.

Für das Modell 4 wurden die Prädiktoren *Geschlecht* und die letzte *Note* (aus dem Regelunterricht Physik) hinzugenommen. Damit verliert die Variable *Muttersprache* an Signifikanz. Dies kann so erklärt werden, dass die Muttersprache und die (Physik-) Note stark korreliert sind und die Note eine Konsequenz der Muttersprache ist. Das wird durch Befunde der PISA-Studie bestätigt (Baumert & Schümer, 2001). Es moderiert daher die Variable *Muttersprache* die Variable Note und nicht umgekehrt. Die Muttersprache ist somit der leistungsfähigere Prädiktor und bereits in Modell 3 berücksichtigt. Das spricht dafür, dass dieses Modell dem Modell 4 vorzuziehen ist. Die Hinzunahme weiterer Prädiktoren zusätzlich zu jenen, die bereits in Modell 3 berücksichtigt sind, ist nicht notwendig. Außerdem wären beide hinzugefügten Prädiktoren nicht signifikant. Das zeigt ebenso die Berechnung des Determinationskoeffizienten (korrigiertes R^2), der 0,253 für Modell 4 beträgt, was sogar geringfügig kleiner ist, als der für Modell 3. Das bedeutet, dass zusätzliche Variable zu keiner akkurateren Modellierung der Posttest-Ergebnisse führen und somit nicht notwendig sind.

Wirklich interessant an Modell 4 ist jedoch, dass die Variable *Geschlecht* sich in dieser Studie als nicht signifikant erwiesen hat. Die fehlende Signifikanz legt nahe, dass mit CAPT eine Unterrichtsform für das Fach Physik gefunden werden konnte, die so gut wie geschlechtsunabhängig wirkt. Das ist insofern erstaunlich, weil es sich hier um ein naturwissenschaftliches Fach handelt, bei denen die Interessen der Mädchen traditionell hinter jenen der Burschen hinterherhinken (z.B. Häußler, et al., 1998). Zwar ist das standardisierte Beta, also der Beitrag, den das Geschlecht liefert, mit -0,048 klein und negativ. Das bedeutet bei der gewählten Kodierung mit 1=weiblich, dass Mädchen ein wenig schlechter anschneiden. Aber eben nur ein wenig, da der Absolutwert dieses Prädiktors sehr klein ist. Obendrein sind wegen der fehlenden Signifikanz keine Unterschiede in der Population zu erwarten. Ein wenig Vorsicht ist mit dieser Interpretation dennoch geboten, da im getesteten Sample die Jungen überrepräsentiert waren, was ebenfalls zu einer sinkenden Signifikanz der Variable *Geschlecht* beitragen kann. In anderen Worten: Aufgrund der größeren Anzahl an Jungen im Sample kann es

sein, dass die Übertretungswahrscheinlichkeit der Variable *Geschlecht* zu hoch geschätzt wurde. Allerdings ist die Übertretungswahrscheinlichkeit mit 0,517 so weit weg von einer signifikanten Größenordnung, dass auch für ein gleichverteiltes Sample nicht von Signifikanz auszugehen ist.

Die Variable *Letzte Note* (in Physik) kann ebenfalls nur wenig zum Modell beitragen. Das ist insofern nicht sehr verwunderlich, da die Schulnoten nicht reliabel sind, zumal sie unstandardisiert von der Lehrkraft vergeben werden. Sie stellen vielmehr eine Rangordnung innerhalb der Klasse dar. Zudem haben innerhalb dieses Samples etwa 70 % der Schüler/innen die Note „Sehr gut“ oder „Gut“.

Ein möglicher Prädiktor, der das Abschneiden im Posttest ebenfalls beeinflussen hätte können, sind die unterschiedlichen Schulstufen der beteiligten Schüler/innen. Dieser Prädiktor ist zwar nicht aus der Theorie abgeleitet, dennoch wurde ein MLR darauf durchgeführt, um mögliche Zweifel auszuräumen, dass die Schulstufe und nicht die Tutorenrolle entscheidend ist. Es zeigte sich jedoch, dass die Schulstufe kein signifikanter Prädiktor für die Posttest-Ergebnisse ist und die aufgeklärte Varianz auch nicht verbessern kann.

Schließlich ermöglicht das MLR-Modell, hier Modell 3, die Schätzung eines Abschneidens verschiedener (fiktiver) Schüler/innen im Posttest, abhängig von den drei Prädiktoren (Tabelle 8.1). Fall 1 beschreibt ein Tutee, das im Praetest lediglich einen Punkt erreicht hatte und Muttersprache Deutsch hat. Der geschätzte Posttestscore beträgt 3,4 Punkte. Fall 2 und Fall 3 unterscheiden sich lediglich in der Rolle. Für eine Schüler/in in der Tutorenrolle wird ein Posttestscore von 8,1 geschätzt, für jemanden in der Tuteerolle lediglich von 5,7.

Fall	Punkte im Praetest	Ist Tutor	Muttersprache	geschätzter Posttestscore
1	1	0	1	**3,4**
2	7	1	1	**8,1**
3	7	0	1	**5,7**
4	1	1	1	**5,8**
5	1	1	0	**5,0**
6	7	1	0	**7,2**

Tabelle 8.1: Schätzung der Posttestscores basierend auf Modell 3 für unterschiedliche Prädiktoren.

Die Fälle 3 und 4 zeigen, dass für sehr unterschiedliche Praetest-Ergebnisse (1 Punkt für einen Tutor vs. 7 Punkte für ein Tutee) die aktive Rolle diesen Unterschied im Posttest fast verschwinden lässt.

Vergleicht man Fall 2 mit Fall 6, so zeigt sich, dass bei gleichen Voraussetzungen (7 Punkte im Praetest) und gleicher Rolle (Tutor) die Muttersprache einen Unterschied von fast einem Punkt (0,9 Punkte) im Posttest erwarten lässt. In Fall 5 und Fall 6 sind sehr unterschiedliche Praetest-Ergebnisse dargestellt. Es zeigt sich hier, dass ein schwächerer Tutor viel aufholen kann gegenüber einem von vornherein guten Tutor.

Im Sinne einer Evaluation einer Methode, hier des CAPT, ist es nötig, sich auch Gedanken über die Persistenz der Intervention zu machen. Die dazu nötige Fragestellung wurde als Forschungsfrage 5 formuliert.

Die Ergebnisse dieser Analysen sind insofern zufrieden stellend, als dass für alle drei Wissenstests der Follow-up Test besser als der Praetest ausfällt. Im Falle der Elektrizitätslehre und des Themas Schatten fällt diese nachhaltige Steigerung sogar hochsignifikant aus und ist mit einer ausreichenden Effektstärke abgesichert. Für das Thema Spiegel ist der Praetest - Follow-up Test Vergleich zwar nicht mehr hochsignifikant. Jedoch ist für alle drei getesteten Bereiche das Vergessen, das im Sinne eines Wissensrückgange als Unterschied zwischen Posttest und Follow-up Test interpretiert wurde, nicht hochsignifikant (vgl. Tabelle 7.40). Das lässt den Schluss zu, dass CAPT in allen getesteten Bereichen zu Verbesserungen im Wissensstand führt und dass nach einigen Wochen dieses Wissen im Sample nicht so weit wieder zurückgegangen ist, dass keine Inferenzen mehr möglich sind. Ein derartiger Effekt ist daher auch in der Population zu erwarten.

Im Rahmen dieser Analysen kann nicht schlüssig interpretiert werden, warum für das Thema Schatten der mittlere Score der Follow-up Tests noch über jenem des Posttests liegt. Über die möglichen Gründe können lediglich Vermutungen angestellt werden. Einerseits kann es sein, dass sich der volle Umfang des Wissens erst nach einer gewissen Reflexionsphase einstellt. Andererseits könnte es sich hier um Test-Retest-Effekte handeln. Beide Erklärungsoptionen abzuklären wäre im Rahmen weiterführender Untersuchungen lohnend.

Die Analysen zu der Langzeitwirkung von CAPT, in Abhängigkeit von den unterschiedlichen Rollen, die die Schüler/innen während der Intervention inne hatten, wurden anhand der Daten zur Elektrizitätslehre durchgeführt. Sie zeigen ein ähnliches Bild: Posttest - Follow-up Test Vergleiche ergeben in keinem Fall eine hochsignifikante Verschlechterung, in manchen aber eine auf einem 5 %-Niveau. Die Gründe, warum ausgerechnet die Tutees, die die kürzeste Interventionszeit erleben durften, den wenigsten Rückgang zu verbuchen haben, sind hier nicht klar. Vermutet werden kann, dass, wie in der Literatur beschrieben, die Interventionszeit auch in unserem Falle keinen Einfluss auf das Ergebnis hat, wenn die Qualität der Instruktion geeignet ist. Da im Falle der Tutees die Posttest-Ergebnisse nicht hochsignifikant über den Praetest-Ergebnissen lagen, kann jedoch ebenso vermutet werden, dass wenig Steigerung im Posttest zu einem geringen Nachlassen im Follow-up Test führt. Beide Vermutungen zu verifizieren sprengt jedoch den Rahmen der hier angestellten Analysen.

Fasst man die Rollen *Tutor* und *Tutees/Tutor* zur *aktiven Rolle* zusammen, ergeben sich im Vergleich von Praetest und Follow-up Test noch immer hochsignifikante Verbesserungen. Dabei tragen zu dieser hochsignifikanten Verbesserung mehrheitlich die Tutees/Tutoren bei, also die Schüler/innen in der Doppelrolle, die eben auch eine größere Interventionszeit hatten. Nachdem zwischen den beiden Gruppen in den Varianzanalysen zu den Posttests keine signifikanten Unterschiede festgestellt werden konnten, kann man dieses Ergebnis dahin gehend interpretieren, dass die längere Interventionszeit zwar die Posttest-Ergebnisse nicht wesentlich beeinflusst, jedoch zur Nachhaltigkeit des Wissen beiträgt.

Nachdem mit Beantwortung der Forschungsfragen 1 bis 5 anhand der Elektrizitätslehre, im Falle von Forschungsfrage 5 auch für die Optik, gezeigt werden konnte, dass CAPT nachhaltige Lernfortschritte bei allen beteiligten Schüler/innen initiiert, wurde für die Forschungsfrage 6 exemplarisch und ausschließlich anhand zweier Bereiche aus der Optik geklärt, in wie weit sich diese Methode auch auf andere Gebiete der Physik erweitern lässt. Dabei war von Beginn an klar, dass für klassenspezifisch unterschiedliche Themen, wie es aus Gründen der Lehrplanvalidität notwendig war, eine Paralleluntersuchung zur Elektrizitätslehre nicht realisierbar war.

Die für die Interventionen zu den Themen Schatten und Spiegel verwendeten Testitems waren in Akzeptanzbefragungen bereits beforscht und adaptiert worden (Haagen-Schützenhöfer & Hopf, 2014) und fußten somit bereits auf Ergebnissen empirischer fachdidaktischer Forschung. Dennoch stellte sich heraus, dass die sprachliche Schwierigkeit dieser Items deutlich über jener der Items zur Elektrizitätslehre lag. Darüber hinaus stellten sie auch wesentlich höhere Anforderungen an das räumliche Vorstellungsvermögen, als die zweidimensionalen Darstellungen in der Elektrizitätslehre. Betrachtet man Abbildung 7.14, so erkennt man, dass die beteiligten Klassen sich eklatant im Anteil der Schüler/innen mit nicht-deutscher Muttersprache unterschieden. Da das auch Einfluss auf den Wissenserwerb in allen anderen Bereichen hat (Baumert & Schümer, 2001) ist es nicht verwunderlich, dass die Klassen sehr unterschiedlich abschnitten. Die bereits in den Praetests festgestellten hochsignifikanten Unterschiede zwischen den Klassen beruhten eventuell auf der sprachlichen Schwierigkeit der Items (Haagen-Schützenhöfer & Hopf, 2014). Da die Analysen zur Elektrizitätslehre essentiell darauf fußten, dass die Praetests in den getesteten Parametern homogen waren, mussten im Bereich der Optik schon aus diesem Grund andere statistische Wege beschritten werden und andere Fragestellungen beantwortet werden, wie eben die nach der Klassenzugehörigkeit und damit nach einem möglichen Zusammenhang zwischen Muttersprache und dem Profitieren von CAPT. Die unterschiedlichen Rollen der Klassen im Tutoringprozess (vgl. Abbildung 7.16) sollten dabei ebenfalls im Auge behalten werden.

Nachdem Prae-Post-Vergleiche über das gesamte Sample signifikante Verbesserungen der Schüler/innen sowohl für das Thema Schatten, als auch für das Thema Spiegel ergeben hatten, sieht es auf Klassenebene etwas anders aus. Die durchgeführte Interaktionsanalyse berücksichtigte zwei Haupteffekte: Zum einen wurde analysiert, ob es Unterschiede in den Gruppenmittelwerten der einzelnen Klassen zwischen den Praetests und den Posttests gibt. Zum anderen wurde analysiert, ob der Testzeitpunkt einen Einfluss auf die Testergebnisse hat, also ob die Schüler/innen im Posttest anders (erwartet wurde besser) abgeschnitten hatten als im Praetest. Das Entscheidende ist jedoch, ob diese Unterschiede durch das Treatment erhalten bleiben oder verändert werden. Diese Frage wurde mit Hilfe von Wechselwirkungsdiagrammen entschieden.

Für die Ergebnisse zum Thema Schatten zeigte sich, dass der Wissenszuwachs zwar hochsignifikant war und von praktischer Bedeutung, mit einer Teststärke von 0,937, aber dass sich auch die Klassen hochsignifikant unterschieden, ebenfalls mit einer hohen Teststärke (0,986). Wird Testzeitpunkt und Klasse gemeinsam getestet, also die Wechselwirkung, welche Klasse wann wie abschneidet, ergeben sich keine signifikanten Befunde und moderate Effektstärken (0,25) bei einem geringen partiellen η^2 (0,045). Für die Interpretation dieser Wechselwirkung ist nun trotz des geringen partiellen η^2 die Beurteilung der Interaktionsdiagramme von Bedeutung. Ohne diese lässt sich auf Basis des festgelegten Signifikanzniveaus lediglich sagen, dass jede der Klassen vom CAPT Treatment profitiert, aber nicht wie sehr. Das ist trotzdem ein schöner Befund, auch wenn die Klassen unterschiedlich waren. Es handelte sich hier um die Klassen 1b, 3, 4, 7 und 9. Diese Klassen sind einander ziemlich ähnlich was die Rolle im CAPT (aktive Rolle, daher Tutoren- oder Tutee/Tutoren- Klassen), die Schulstufe (7. und 8.) und Schülerzahl betrifft. Auf den ersten Blick liegt der Unterschied im Anteil an Schüler/innen mit nicht-deutscher Muttersprache. Dieser wurde daher auch näher betrachtet, auch weil die Ergebnisse aus der MLR zur Elektrizitätslehre das nahe legen. Klasse 3 ist jene mit dem höchsten Anteil an Schüler/innen mit nicht-deutscher Muttersprache (über 80 %), während diese Anteile in den Klassen 4 und 7 unter 20 % liegen. Alle Klassen verbessern sich im Posttest, wobei Klasse 3 nicht den geringsten Zuwachs erzielte. Ein hoher Anteil an nicht muttersprachlich deutschen Kindern bedeutet somit nicht, dass CAPT nicht wirkt. Im Gegenteil, CAPT scheint jedenfalls zu wirken. Das wäre vielleicht aufgrund der Ergebnisse aus der Elektrizitätslehre nicht zu erwarten gewesen, da in der MLR die Muttersprache ein entscheidender Prädiktor für den Posttestscore ist. Die Folgerung ist, dass CAPT zwar besser wirkt, wenn die Unterrichtssprache und die gesprochene Sprache identisch sind, dass aber der Umkehrschluss nicht zulässig ist, nämlich, dass CAPT nicht wirkt, wenn dem nicht so ist. Ganz im Gegenteil: Betrachtet man die Darstellung des Haupteffektes Klasse (Abbildung 7.18), so zeigt sich, dass aufgrund der nicht erhaltenen Rangordnung kein eindeutiger Schluss von den Klassen auf den Wissenszuwachs (Differenz Posttest – Praetest) möglich ist.

Ein Beispiel kann das illustrieren: Klasse 3 zeigt eine stärkere Verbesserung als Klasse 1 und eine vergleichbare wie Klasse 9. Nun haben aber diese Klassen sehr unterschiedliche

Anteile an Schüler/innen mit nicht-deutscher Muttersprache. Klasse 3 hat mit Abstand den höchsten Anteil, Klasse 1 nur etwa die Hälfte und Klasse 9 wiederum die Hälfte davon (Abbildung 7.14). Die Klassenzugehörigkeit und damit der Anteil an Schüler/innen mit nicht-deutscher Muttersprache lassen somit keine Prognose für die Verbesserung im Posttest zu.

Das ist nach den Ergebnissen der MLR zur Modellierung der Posttest-Ergebnisse in der Elektrizitätslehre überraschend. Insbesondere, da die Textschwierigkeit der Testitems zum Thema Optik deutlich höher ist und schon zum Verstehen der Fragestellung sinnerfassendes Lesen nötig ist. Offensichtlich erzeugt diese Textschwierigkeit zwar große Unterschiede im Ausgangsniveau wie auch im Endergebnis, lässt aber keine Schlüsse über die Verbesserung der einzelnen Schüler/innen und damit über die Wirksamkeit von CAPT auf unterschiedliche Schülergruppen zu. Man kann vorsichtig daraus schließen, dass CAPT unabhängig vom Sprachniveau zumindest die der Gruppe der aktiven Schüler/innen jedenfalls zu Fortschritten führt. Es scheint jedoch nicht dazu geeignet zu sein, Unterschiede aufgrund der sprachlichen Ausgangslage zu beseitigen.

Betrachtet man nun die Ergebnisse zum Thema Spiegel zeigt sich, dass hier der Wissenszuwachs ebenfalls hochsignifikant ist, mit einer Teststärke von 0,990 und dass sich auch die Klassen hochsignifikant und mit einer hohen Teststärke (0,859) voneinander unterscheiden. Testzeitpunkt und Klasse ergeben, wie beim Thema Schatten, keine signifikante Wechselwirkung bei einer ebenfalls moderaten Teststärke (0,269) und einem geringen partiellen η^2 (0,058). Auch hier wurden wiederum die Wechselwirkungsdiagramme betrachtet um den Effekt stichhaltig interpretieren zu können. Im Unterschied zum Thema Schatten zeigt sich beim Spiegel, dass beide Haupteffekte interpretierbar sind, da für beide Interaktionsdiagramme die jeweiligen Rangordnungen erhalten bleiben. Der Posttest fiel hochsignifikant besser aus, als der Praetest und die Reihenfolge der Klassen blieb erhalten. Das bedeutet, dass sich alle Klassen verbessern, und man vorhersagen kann, dass im Praetest schlechtere Klassen auch im Posttest schlechter abschneiden werden, hingegen im Praetest bessere Klassen auch im Posttest besser abschneiden werden. Darüber hinaus lässt sich ablesen, dass jene Klasse, die im Praetest am schlechtesten abgeschnitten hatte (Klasse 2), den wenigsten Wissenszuwachs verzeichnen konnte. Die Klasse (Klasse 10), die im Praetest

am besten war, zeigte auch den größten Wissenszuwachs. Aufgrund der Interpretierbarkeit des Haupteffektes Klassenzugehörigkeit, der noch dazu hochsignifikant war, lässt sich ein derartiges Erscheinungsbild auch in der Population vermuten.

Interessant ist jedoch, dass diese drei Klassen unterschiedliche Rollen innerhalb von CAPT belegten: So war Klasse 2 eine Tutorenklasse, Klasse 8 eine Tutee-Klasse und die Schüler/innen der Klasse 10 belegten beide Rollen. Trotzdem zeigte die Tutee-Klasse einen deutlich besseren Posttest und auch einen größeren Wissenszuwachs als Klasse 2, die Tutoren waren.

Die Befundlage zu den Themen Schatten und Spiegel ist daher alles andere als eindeutig. Zeigt sich beim Thema Schatten, dass die Klassenzugehörigkeit keinen Rückschluss auf ein Abschneiden im Posttest zulässt, ist das beim Thema Spiegel sehr wohl der Fall. Beim Thema Schatten waren alle Klassen in der aktiven Rolle. Beim Thema Spiegel hatten unterschiedliche Klassen unterschiedliche Rollen inne. Jedoch ist hier eine Reihung in dem Sinne nicht möglich, dass die Klassen in der aktiven Rolle besser abschnitten. Eine Interpretation des Beitrags der aktiven Rolle ist im Rahmen dieser beiden Optik-Themen somit nicht möglich.

Versucht man die Unterschiede in den Klassen am Anteil der Schüler/innen mit nicht-deutscher Muttersprache festzumachen zeigt sich für beide Themen, dass jene Klassen am schlechtesten abschneiden, die mit über 80 % den höchsten Anteil daran haben (Klassen 2 und 3). Die besten Ergebnisse, sowohl im Posttest, als auch im Wissenszuwachs, zeigen jene Klassen, die den geringsten Anteil an Schüler/innen mit nicht-deutscher Muttersprache haben. Das sind die Klassen 4, 7, 8 und 10 mit jeweils weniger als 20 %.

Ein möglicher Grund dafür, dass es beim Thema Schatten keinen eindeutigen Effekt der Klassenzugehörigkeit gibt, aber beim Thema Spiegel schon, mag in der unterschiedlichen Schwierigkeit der Themen liegen. Das Thema Schatten war für Interventionen mit jüngeren Tutees gedacht, das Thema Spiegel für ältere. Daher ergab es sich, dass die Interventionen auf unterschiedlichem Niveau abliefen und auch die Testitems für den Schatten, obgleich schwieriger als in der Elektrizitätslehre, etwas einfacher und vor allem weniger abstrakt waren als jene für den Spiegel. Man kann daher für beide

Themen von unterschiedlichen Schwierigkeitsgraden ausgehen, für die Tutees, wie auch für die Tutoren. Hierbei ist es durchaus möglich, dass beim schwierigeren, abstrakteren Thema Spiegel, wo mehr erklärt werden musste und weniger gezeigt werden konnte, die Sprachbeherrschung eine wesentlichere Rolle spielte als beim Schattenthema. Daher konnten beim Schattenthema auch sprachlich schwächere Klassen erheblich profitieren, was beim Spiegelthema auf Limits stieß.

Trotzdem muss festgestellt werden, dass die beiden Klassen, Klasse 2 (Spiegel) und 3 (Schatten), die jeweils einen Anteil von über 80 % an Schüler/innen mit nicht-deutscher Muttersprache hatten, diejenigen sind, die sowohl in den Praetests, wie auch in den Posttests mit Abstand am schlechtesten abschnitten. Vom Wissens*zuwachs* her ist allerdings Klasse 3 im Bereich des Schattens nicht die schlechteste, was wiederum mit der relativen Einfachheit des Themas in Zusammenhang gebracht werden kann. Klasse 2 hingegen zeigt die geringste Verbesserung.

Fasst man die Befunde zu den Wissenstests aus der Optik mit denen aus der Elektrizitätslehr zusammen, kann man davon ausgehen, dass CAPT jedenfalls zu einem signifikant verbesserten Posttest-Ergebnis führt. Eine Abhängigkeit vom Praetestscore und der Muttersprache der Schüler/innen scheint wahrscheinlich. Es drängt sich die Vermutung auf, dass CAPT in Abhängigkeit von der Schwierigkeit des Themas erst ab einem gewissen Mindestniveau der Klassen wirkt.

Diese Fragestellungen konnten im Rahmen der hier präsentierten Arbeit nicht hinreichend genau beantwortet werden. Ihnen noch einmal genauer nachzugehen wäre sicher ein lohnenswertes Forschungsobjekt für eine Folgestudie. Insbesondere wäre es wünschenswert herauszufinden, ob sich der Unterschied zwischen den Klassen, wie er eben ersichtlich war, tatsächlich aus dem Anteil an Schüler/innen mit nicht-deutscher Muttersprache ergibt oder ob hier noch andere Faktoren beteiligt sind, wie z.B. das kognitive Niveau oder Einstellungen zum Lernen.

Die Zusammenhänge zwischen den getesteten nicht-kognitiven Parametern und den Testergebnissen waren in fast allen Fällen nicht aussagekräftig. Es kann durchaus als enttäuschend gewertet werden, dass die Motivationstests sich nicht als Prädiktoren für Posttest-Ergebnisse heranziehen ließen, einerseits wegen der fehlenden Signifikanz im Bereich der Optik, andererseits wegen der geringen Stärke des Zusammenhanges in der

Elektrizitätslehre. In den getesteten multilinearen Regressionsmodellen zu den Posttest-Ergebnissen der Elektrizitätslehre war der Prädiktor Motivation, egal welcher Nuance und mit welchem Instrument gemessen, niemals signifikant, was die Ergebnisse aus den Korrelationsanalysen bestätigt.

Das selbst entwickelte Motivationsinstrument kann durchaus als noch nicht ausgereift erachtet werden, da für die restlichen Skalen, außer der *effort/importance* Skala, noch einiges an Entwicklungsarbeit für treffsichere Items offen ist. Daher war es auch nicht überraschend, dass sich nur schwache Zusammenhänge zu den Wissenstests ergeben hatten. Allerdings ergeben sich mittelstarke und teilweise signifikante Korrelationen zu den Ergebnissen der restlichen Motivationstests, die parallel dazu erhoben wurden (FAM, SDI, SWE, vgl. Kapitel 5.3), und die zwar nicht dieselben, aber teilweise überlappende Konstrukte abbildeten. Das kann als Zeichen für die Konstruktvalidität des Motivationsinstrumentes erachtet werden. Dennoch lassen diese psychometrisch validen Testinstrumente ebenso keinen aussagkräftigen Zusammenhang zu den Ergebnissen der Wissenstests erkennen. Über mögliche Gründe kann man an dieser Stelle nur spekulieren. Das kann einerseits daran liegen, dass diese Instrumente alle sehr feine Nuancen abfragen, die nicht einem, möglicher Weise weniger feinen, Reflexionsgrad der Schüler/innen entsprechen. Andererseits kann es auch sein, dass Schüler/innen dieser Altersstufe noch nicht über die nötige Reife verfügen, die intrinsische Motivation ermöglicht (vgl. Kapitel 2.3).

Da es sich bei CAPT um eine Unterrichtsmethode handelt, ist neben allem akademischen Interesse nicht zuletzt die Praxistauglichkeit entscheidend, ob eine Implementierung in den Schulalltag gelingt. Aspekte der Praxistauglichkeit wurden bereits bei der Gestaltung des Forschungsdesigns im Auge behalten, indem zum einen ganze Klassen miteinander arbeiteten, die Tutor-Tutee-Zuordnung nicht explizit gesteuert wurde und der Zeitrahmen für das eigentliche Tutoring etwa eine Unterrichtseinheit ausmachte. Eine derartige Vorgehensweise ist auch in der Schule leicht und ressourcenschonend möglich, da CAPT sich in bestehende Stundenpläne einfügen lässt und die Auswahl und Zuordnung der Schüler/innen zueinander ohne a priori diagnostischen Aufwand möglich ist. Zudem zeigte sich, dass aufgrund der positiven Wirkung von CAPT auf die Tutoren

zusätzliche Belohnungsprogramme nicht von Priorität sind um deren Aufwand zu entschädigen.

Trotz dieser vielleicht wenig differenziert erscheinenden, aber praktisch leicht durchführbaren Vorgangsweise, zeigen die Forschungsergebnisse klare Erfolge der Methode. Darüber hinaus versteht sich CAPT als konstruktivistische Lernumgebung, die die aktive Konstruktion von Konzepten unterstützt, falls man gewisse Empfehlungen in der Durchführung einhält. Dazu zählen die Orientierung an Schülervorstellungen ebenso, wie die P-O-E Strategie bei der Durchführung von Experimenten.

Aus diesen Gründen kann CAPT ohne Einschränkungen als ergänzende Methode im schulischen Regelunterricht empfohlen werden, auch wenn einige Mechanismen der konzeptuellen Entwicklung und der motivationalen Beeinflussung mit dieser Arbeit nicht restlos geklärt werden konnten.

9. Zusammenfassung und Ausblick

Im Rahmen der vorliegenden empirischen Arbeit wurde Cross-Age Peer Tutoring in den hier vorgestellten physikalischen Inhaltsbereichen, anhand der vorgestellten Stichprobe aus der Sekundarstufe 1 evaluiert. Als Kriterien ergaben sich zwei Aspekte: die Wissensentwicklung und die Motivation der Schüler/innen.

CAPT versteht sich hierbei als konstruktivistischer Unterrichtsansatz, der durch die beiden Phasen Mentoring und Tutoring die zur Eigenkonstruktion von Konzepten nötigen Lernumgebungen bieten soll. Innerhalb der Gestaltung dieser Lernumgebungen wurde die Orientierung am Vorwissen und an Schülervorstellungen als zentraler Aspekt der inhaltlichen Fokussierung gesehen. Die Möglichkeit der Interaktion stellte einen wesentlichen motivationalen Aspekt dar und gab Gelegenheit zu Aushandlungs-prozessen unter Schüler/innen.

Das Mentoring diente den Tutoren als Vorbereitung, sowohl in inhaltlicher, wie auch in methodischer Hinsicht und dauerte etwa 60 - 80 Minuten. Die inhaltliche Auseinander-setzung basierte auf der Gelegenheit zur Reflexion der eigenen Vorstellungen anhand von unterschiedlichen experimentellen wie nicht-experimentellen Aufgaben. Darüber hinaus sollten die Tutoren für die Vorstellungen der Tutees sensibilisiert werden. Methodisch wurde die P-O-E Strategie (vgl. Kapitel 4.4) für die Arbeit mit den Tutees vorgeschlagen. Das eigentliche Tutoring fand ein bis zwei Wochen danach statt und dauerte, anschließend an eine etwa 20-minütige Wiederholung für die Tutoren, weitere 30 - 45 Minuten.

Im gesamten CAPT Projekt wurden in zwei Durchgängen 172 bzw. 141 Lernende zu Themen aus der Elektrizitätslehre und der Optik beforscht. Sie konnten im Peer Tutoring Prozess unterschiedliche Rollen inne haben: als Tutoren, die aktiv am Geschehen beteiligt waren, indem sie zur inhaltlichen Gestaltung beitragen konnten, oder aber als Tutees, die betreut wurden. Diese bekleideten die mehrheitlich passive Rolle, in der sie aber dennoch selbstständig experimentell agieren konnten, jedoch keinen Einfluss auf die Inhalte hatten. Daneben gab es auch Schüler/innen, die reziprokes Tutoring erfahren durften, indem sie nacheinander beide Rollen bekleideten.

Die erste und grundlegendste Fragestellung war jene, ob CAPT überhaupt im Kontext von Physik in der beforschten Altersstufe Erfolge zeigt. Die beobachteten Effektstärken zwischen 0,46 und 0,62 in den Wissenstests, sowohl im Bereich der Elektrizitätslehre als auch im Bereich der Optik, lassen diese Frage mit einem klaren „Ja“ beantworten. Diese Verbesserung in den Testergebnissen kann als Hinweis darauf gesehen werden, dass nach der CAPT-Intervention belastbarere Konzepte vorliegen und somit ein Konzeptwechsel unterstützt wurde.

Zur Frage nach der motivationalen Wirksamkeit von CAPT kann keine empirisch gesicherte Antwort gegeben werden, die über die Einschätzungen der betreffenden Klassenlehrer/innen hinaus geht. Sowohl das Messinstrument zur Motivation, das im Rahmen dieser Arbeit aus dem IMI adaptiert und zum Teil neu entwickelt wurde, als auch bereits bestehende, auf ihre Validität und Reliabilität getesteten Instrumente wie der FAM, der SDI oder das Messinstrument zur Selbstwirksamkeitserwartung (vgl. 5.3), offenbarten keine Zusammenhänge, wie sie aufgrund theoretischer Überlegungen zu erwarten gewesen wären.

Eine detailliertere Analyse der Posttest-Ergebnisse, hinsichtlich eines möglichen Einflusses der Rollen, wurde im Rahmen der Elektrizitätslehre durchgeführt. Es zeigte sich, dass jene Schüler/innen, die zumindest einmal in der Rolle der Tutoren waren und daher aktiv am Geschehen teilnahmen, im besonderen Maße von CAPT profitierten. Dabei waren hier keine Dosiseffekte zwischen den Tutoren und den Tutee/Tutoren nachzuweisen, die unterschiedlich lange Interventionen erlebten. Nachdem die Praetests über das gesamte Sample hinweg sehr homogen waren, ließ sich schließen, dass die gefundenen Unterschiede in den Posttests auf die Rolle im Tutoring Prozess zurück zu führen waren.

Multiple lineare Regressionsmodelle, die die Ergebnisse der Analysen zu den Posttests zusammenfassten und gleichzeitig einen Vergleich der zu erwartenden Stärke der einzelnen Effekte in der Population erlaubten, lieferten folgende signifikante und bedeutsame Prädiktoren für die Posttest-Ergebnisse: die Praetest-Ergebnisse, die das Ausgangsniveau beschreiben; die aktive Rolle; und die Muttersprache, wobei mit dieser Variablenbezeichnung gemeint ist, ob die Sprache der Instruktion gleich der zu Hause gesprochenen Sprache ist.

Eine interessante Folgerung aus den verschiedenen Regressionsmodellen ist, dass der Prädiktor *Geschlecht*, ebenso wir die *Note* im Regelunterricht, keinen signifikanten Einfluss auf die Posttest-Ergebnisse hat. Ist dieses Ergebnis für die Variable Note nicht überraschend, da *Noten* nicht-standardisiert vergeben werden, so überrascht es doch hinsichtlich des *Geschlechts*. Es scheint so zu sein, dass mit CAPT als Unterrichtsmethode ein Zugang identifiziert werden konnte, der das bestehende Ungleichgewicht zwischen Jungen und Mädchen hinsichtlich des Interesses und der Leistungen im Physikunterricht zumindest nicht verstärkt.

Für die Themen der Optik wurden Analysen auf Klassenebene durchgeführt. Aufgrund der sprachlich schwierigeren Testitems wurde der Grund für unterschiedliches Abschneiden in den Posttests im differierenden Anteil von Schüler/innen mit nicht-deutscher Muttersprache vermutet und diese Hypothese auch untersucht.

In Interaktionsanalysen mit den Haupteffekten Klasse und Testzeitpunkt konnte gezeigt werden, dass die Posttests jedenfalls hochsignifikant besser ausfielen als die Praetests. Die Untersuchungen zur Muttersprache zeigten, dass CAPT auch in Klassen mit einem hohen Anteil an Schüler/innen mit nicht-deutscher Muttersprache wirkt, jedoch nicht dazu geeignet ist, bereits bestehende Unterschiede im Ausgangsniveau zu reduzieren. Eindeutige Zuordnungen zwischen einem Anteil an Schüler/innen mit nicht-deutscher Muttersprache und Erfolgen bei CAPT waren jedoch nicht möglich. Die Befunde können aber dahin gehend gedeutet werden, dass mit steigender sprachlicher Schwierigkeit der Items und größerer Abstraktheit der fachlichen Inhalte, der Muttersprache immer mehr Bedeutung zukommt, und CAPT eventuell erst ab einem gewissen kognitiven oder sprachlichen Mindestniveau wirkt.

Für Evaluationen ist die Persistenz der untersuchten Intervention ebenso interessant wie die aktuelle Wirkung. Im Rahmen von Analysen der Follow-up Tests konnte zumindest gezeigt werden, dass zwei bis fünf Wochen nach der Intervention die Testergebnisse in allen getesteten Bereichen immer noch besser ausfielen als die Praetests. Während diese nachhaltigen Steigerungen nur für die Elektrizitätslehre und das Thema Schatten signifikant bleiben, ist der Unterschied zwischen Posttest und Follow-up Test, der als Maß für das Vergessen interpretiert wird, für kein Thema der Intervention signifikant. Untersuchungen hinsichtlich der Rolle zeigten, dass die Schüler/innen in der Doppelrolle

(Tutees/Tutoren) die nachhaltigste Steigerung erzielten, was hier vielleicht mit der Dauer der Intervention in Zusammenhang gebracht werden kann. Summa summarum kann somit gefolgert werden, dass CAPT zu einem Wissenszuwachs führt, der über die unmittelbare Dauer der Intervention hinaus bestehen bleibt.

Die praktische Bedeutung von CAPT lässt sich aus dem Vergleich zwischen Erfolg der Unterrichtsmethode und dem Aufwand, mit dem sich diese Methode in den Schulalltag implementieren lässt, bewerten. Hier ist einerseits von Bedeutung, dass alle Schüler/innen, insbesondere die Tutoren, von CAPT profitieren. Daher ist es nicht nötig, spezielle Belohnungsprogramme für die Tutoren zu entwickeln. Die Belohnung ist der Fortschritt im Fach selbst. Andererseits sind spezielle Diagnosen zur Auswahl der Tutoren z.B. nach Können, nicht nötig. CAPT ist für eine breit gestreute Schülerpopulation geeignet, und seine Wirksamkeit beschränkt sich nicht nur auf *High-Achiever*. Es ist problemlos möglich, in eins-zu-eins Settings ganze Tutorenklassen mit ganzen Tuteeklassen zu kombinieren. Trotzdem sind nennenswerte Effekte zu erwarten.

Anschließend an die Ergebnisse dieser Arbeit ergeben sich einige offene Punkte, die in diesem Rahmen nicht vollständig geklärt wurden und die Thema möglicher Folgestudien sein können. Die sicherlich interessanteste Fragestellung ist, welchen Einfluss CAPT auf die konzeptuelle Struktur der Schüler/innen besitzt. In der vorliegenden Untersuchung finden sich nachhaltige Verbesserungen im Wissen, die als Einleitung zu einem Konzeptwechsel verstanden werden können. Eine endgültige Antwort kann mit rein quantitativen Methoden nicht gegeben werde und ist in Kombination quantitativer und qualitativer Methoden zu erwarten.

Außerdem sind Einflüsse des Mentorings und Test-Retest-Effekte nicht eindeutig geklärt, was mit einem Vergleichsgruppendesign möglich wäre.

Das etwas überraschende Ergebnis, dass CAPT für Mädchen genau so gut wie für Jungen wirkt, wäre sicherlich ein lohnenswerter Untersuchungsgegenstand für eine eigene Studie. Es wäre zu überprüfen, ob dieses Ergebnis nur in der Elektrizitätslehre zu finden ist, oder ob es sich auch auf andere Themen erweitern ließe. Dabei wäre es aus Sicht der Inferenzstatistik sinnvoll, die Untersuchungen an einem Sample durchzuführen, in dem Jungen und Mädchen gleich häufig vertreten sind, um Verzerrungen bei der Schätzung der Übertretungswahrscheinlichkeit (Signifikanz) zu vermeiden.

Eine lohnenswerte Fragestellung könnte sich der, im Rahmen dieser Arbeit aufgeworfenen, aber nicht restlos geklärte Vermutung widmen, dass CAPT erst ab einem gewissen Mindestniveau wirkt, sowie der Klärung der Interferenz von sprachlichen Fähigkeiten, konzeptuellen Schwierigkeiten, Itemschwierigkeit und Rolle im Tutoringprozess. Eine derartige Untersuchung wäre vor dem Hintergrund sozioökonomisch und sprachlich-kulturell heterogener Klassen, wie sie sich in Ballungsräumen vermehrt finden, eine wichtige Angelegenheit, zumal vermutet werden kann, dass CAPT sich für derartige Schülerpopulationen gut eignet.

10. Literatur

Allison, P. D. (2009). Missing Data. In R. E. Millsap (Ed.), *The Sage Handbook of Quantitative Methods in Psychology* (pp. 72-89). Thousand Oaks: Sage.

Andersson, B., & Kärrquist, C. (1983). How Swedish pupils, aged 12-15 years, understand light and its properties. *International Journal of Science Education, 5*, 387-402.

Apostolopoulos, N., & Schulz, A. (2003). Klumpenstichproben. *Projekt Neue Statistik 2003* Retrieved 01.06., 2014, from http://web.neuestatistik.de/inhalte_web/content/files/modul_29064.pdf

Atkinson, J. W. (1975). *Einführung in die Motivationsforschung*. Stuttgart: Klett.

Aufschnaiter, S. v., Fischer, H. E., & Schwedes, H. (1992). Kinder konstruieren Welten. Perspektiven einer konstruktivistischen Physikdidaktik. In S. J. Schmidt (Ed.), *Kognition und Gesellschaft. Diskurs des radikalen Konstruktivismus.* Frankfurt am Main: Suhrkamp.

Bandura, A. (1977). Self-Efficacy - Toward a Unifying Theory of Behavioral Change. *Psychological Review, 84*(2), 191-215.

Bandura, A. (1997). *Self Efficacy: The Exercise of Control* (Vol. Capter 6. Cognitive Functioning). New York: Freeman.

Barab, S., & Squire, K. (2004). Design-Based Research: Putting a Stake in the Ground. *The Journal of the Learning Sciences, 13*(1), 1-14.

Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung udn Kompetenzerwerb. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Eds.), *PISA 2000*. Opladen: Leske + Budrich.

Berger, R. (2000). *Moderne bildgebende Verfahren der medizinischen Diagnostik - Ein Weg zu interessantem Physikunterricht* (Vol. 11). Berlin: Logos.

Berger, R. (2011). Motivationsitems. Persönliche Korrespondenz. Wien.

Berger, R., & Hänze, M. (2004). Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II–Einfluss auf Motivation, Lernen und Leistung. *Zeitschrift für Didaktik der Naturwissenschaften, 10*, 205-219.

BMUKK. (2000). Lehrpläne der AHS Unterstufe. Retrieved 07/22, 2013, from http://www.bmukk.gv.at/schulen/unterricht/lp/lp_ahs_unterstufe.xml

Bortz, J., & Döring, N. (2003). *Forschungsmethoden und Evaluation*. Berlin: Springer.

Breit, S. (2009). Sozialisationsbedingungen von Schülerinnen und Schülern mit Migrationshintergund. In C. Schreiner, Schwantner, U. (Ed.), *PISA 2006* (pp. 136-145). Graz: Leycam.

Brosius, F. (1998). *SPSS 8. Professionelle Ststistik unter Windows*. Heidelberg: mitb.

Bruneforth, M., Herzog-Punzenberger, B., & Lassnigg, L. (2013). *Nationaler Bildungsbericht Österreich 2012: Indikatoren und Themen im Überblick*. Graz: Leykam.

Buehner, M., & Ziegler, M. (2009). *Statistik fuer Psychologen und Sozialwissenschaftler*.

Muenchen: Pearson.

Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion*. München: Pearson.

Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson.

Busker, M. (2014). Entwicklung eines Fragebogens zur Untersuchung des Fachinteresses. In D. Krüger, I. Parchmann & H. Schecker (Eds.), *Methoden in der naturwissenschafts-didaktischen Forschung*. Berlin, Heidelberg: Springer.

Clark, R. E., Kirschner, P. A., & Sweller, J. (2012). Putting Students on the Path to Learning. The Case for Fully Guided Instruction. *AmERIcAN EdUcATOR, 36*(1), 6-11.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.): Lawrence Erlbaum.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational Outcomes of Tutoring - A Meta-Analysis of Findings. *American Educational Research Journal, 19*(2), 237-248.

Colin, P., Chauvet, F., & Viennot, L. (2002). Reading images in optics: students difficulties and teachers views. *International journal of Science Education, 24*(3), 313-332.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive Apprenticeship: Making thinking visible. *Americaqn Educator, 15*(3), 6-11.

Collins, A., Brown, J. S., & Newman, S. E. (Eds.). (1989). *Cognitive Apprenticeship: Teaching the crafts of reading, writing and mathematics.* New York: Hillsdale: Lawrence Erlbaum Associates.

Creswell, J. W., & Garrett, A. L. (2008). The "movement" of mixed methods research and the role of educators. *South African Journal of Education, 28*(3), 321-333.

Cronbach, L. J., & Snow, R. E. (1981). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*: Ardent Media.

Deci, E. L. (2014). Selfdetermination Theory. Persönliche Korrespondenz. Wien.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L., & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39*(2), 223-238.

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry, 11*(4), 227-268.

Deci, E. L., & Ryan, R. M. (2003). Intrinsic Motivation Inventory. Retrieved 12-23, 2013, from http://selfdeterminationtheory.org/edu/scales/category/5-intrinsic-motivation-inventory

Deci, E. L., & Ryan, R. M. (2008). Self-Determination Theory: A Macrotheory of Human Motivation, Development, and Health. *Canadian Psychology, 49*(3), 182-185.

Deci, E. L., & Ryan, R. M. (Eds.). (2002). *An Educational-Psychological Theory of Interest and Its Relation to SDT*. Rochester: University of Rochester Press.

Duit, R. (Ed.). (1999). *Conceptual change approaches in science education*. Amsterdam: Pergamon.

Duit, R., & Rhöneck, C. (1998). Learning and understanding key concepts of electricity. *Connecting research in physics education with teacher education*, 55-62.

Duit, R., & Treagust, D. F. (2003). Conceptual Change: A powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*(6), 671-688.

Duit, R., & Treagust, D. F. (2012). Conceptual change: Still a powerful framework for improving the practice of science instruction. In K. C. T. Tan & M. Kim (Eds.), *Issues and Challenges in Science Education Research* (pp. 43-54). Heidelberg: Springer.

Duit, R., Treagust, D. F., & Widodo, A. (2008). *Teaching Science for Conceptual Change: Theory and Practice*. New York: Taylor & Francis.

Dweck, C. S., & Reppucci, N. D. (1973). Learned helplessness and reinforcement responsibility in children. *Journal of Personality ans Social Psychology, 25*(1), 109-116.

Eckhardt, A. G. (2008). *Sprache als Barriere für den schulischen Erfolg: Potentielle Schwierigkeiten beim Erwerb schulbezogener Sprache für Kinder mit Migrationshintergrund* (Vol. 9). Münster: Waxmann Verlag.

Engelhardt, P. V., & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics, 72*, 98-115.

Ertl, D. (2014). Sechs Kernaspekte zur Natur der Naturwissenschaft. *Plus Lucis, 1-2*, 16-20.

Fogarty, J. L., & Wang, M. C. (1982). An Investigation of the Cross-Age Peer Tutoring Process: Some Implications for Instructional Design and Motivation. *The Elementary School Journal, 82*(5), 451-469.

Freiburg, A.-L.-U. (2009). Tutorat 8 - Wiederholung Faktorenanalyse. Retrieved 05.05., 2014, from http://www.psychologie.uni-freiburg.de/abteilungen/Sozialpsychologie.Methodenlehre/courses/ss-09/spss-und-statistik/tutorat8.ppt/download.

Galili, I., & Hazan, A. (2000). Learners' knowledge in optics: interpretation, structure and analysis. *International Journal of Science Education, 22*(1), 57-88.

Gaustad, J. (1993). *Peer and Cross-Age Tutoring*. Oregon: ERIC Clearinghouse on Educational Management Eugene.

Gerstenmaier, J., & Mandl, H. (1995). Wissenserwerb unter konstruktivistischer Perspektive. *Zeitschrift für Pädagogik, 41*(6), 867-888.

Glasersfeld, E. v. (1997). Kleine Geschichte des Konstruktivismus. *Österreichische Zeitschrift für Geschichtswissenschaft, 8*(1), 9-17.

Goldberg, F. M., & McDermott, L. C. (1986). Student difficulties in understanding image formation by a plane mirror. *The Physics Teacher, 24*(8), 472-480.

Gollwitzer, M., & Jäger, R. S. (2009). *Evaluation kompakt*. Weinheim: Beltz.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Grzega, J., & Schöner, M. (2008). The didactic model LdL (Lernen durch Lehren) as a way of preparing students for communication in a knowledge society. *Journal of Education for Teaching, 34*(3), 167-175.

Guesne, E. (1984). Die Vorstellung von Kindern über Licht. *physica didactica, 11*(2), 3.

Guesne, E. (1985). Light. In R. Driver, E. Guesne & A. Tiberghien (Eds.), *Children's ideas in science.* (pp. 10-32). Buckingham: Open University Press.

Haagen-Schützenhöfer, C. (2011). Testitems zum Thema Spiegel. Persönliche Korrespondenz. Wien.

Haagen-Schützenhöfer, C., & Hopf, M. (2011). *Entwicklung eines Testinstruments zur geometrischen Optik.* Paper presented at the Jahrestagung der GDCP, Oldenburg.

Haagen-Schützenhöfer, C., & Hopf, M. (2014). Development of a two-tier test-instrument for geometrical optics. In C. Constantinou (Eds.), E-Book Proceedings of the ESERA 2014 Conference: Science Education Research for Evidence-based Teaching and Coherent Learning

Haagen-Schützenhöfer, C., Rottensteiner, J., & Hopf, M. (2012). *Akzeptanzbefragung zu Optikunterrichtsmaterialien.* Paper presented at the Jahrestagung der GDCP, Hannover.

Haagen-Schützenhöfer, C., & S., G. (2014). *Assessing Students' Knowledge on Refraction.* Paper presented at the NDSTE 2014, Corfu.

Haider, G., & Reiter, C. (Eds.). (2004). *PISA 2003. Internationaler Vergleich von Schülerleistungen*. Graz: Leykam.

Hattie, J. A. C. (2009). *Visible Learning: A sythesis of over 800 meta-analyses relating to achievement*. London, New York: Routledge.

Häußler, P., Bünder, W., Duit, R., Gräber, W., & Mayer, J. (1998). *Naturwissenschaftsdidaktische Forschung: Perspektiven für die Unterrichtspraxis*. Kiel: Institut für die Pädagogik der Naturwissenschaften.

Helmke, A., & Weinert, F. E. (Eds.). (1997). *Bedingungsfaktoren schulischer Leistungen* (Vol. 3). Göttingen: Hogrefe.

Howe, C., Tolmie, A., Greer, K., & Mackenzie, M. (1995). Peer collaboration and conceptual growth in physics: Task influences on children's understanding of heating and cooling. *Cognition and Instruction, 13*(4), 483-503.

Kim, B. J., & Gill, D. L. (1997). A Cross-Cultural Extension of Goal Perspective Theory to Korean Youth Sport. *Journal of Sport and Exercise Psychology, 19*, 142-155.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels (3rd edition)*. San Francisco: Beterr-Koehler.

Knorr-Cetina, K. (1989). Spielarten des Konstruktivismus. *Soziale Welt, 40*(1/2), 86-96.

Korner, M., & Hopf, M. (2014). Cross-Age Peer Tutoring in Physics: Tutors, Tutees, and Achievement in Electricity. *International Journal of Science and Mathematics Education*. Retrieved from http://dx.doi.org/10.1007/s10763-014-9539-8.

doi:10.1007/s10763-014-9539-8

Korner, M., Urban-Woldron, H., & Hopf, M. (2012). *Entwicklung eines Messinstrumentes zur Motivation.* Paper presented at the GDCP Jahrestagung - Konzepte fachdidaktischer Strukturierung für den Unterricht, Oldenburg.

Krapp, A. (Ed.). (2002). *An Educational-Psychological Theory of Interest and Its Relation to SDT*. Rochester: University of Rochester Press.

Krapp, A., & Ryan, R. M. (2002). Selbstwirksamkeit und Lernmotivation. In M. Jerusalem & D. Hopf (Eds.), *Zeitschrift für Pädagogik. Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen* (Vol. 44, pp. 54-82).

Krüger, D. (2012). Die Conceptual Change-Theorie. In D. Krüger & H. Vogt (Eds.), *Theorien fachdidaktischer Forschung*.

Lamnek, S. (2005). *Qualitative Sozialforschung*. Weinheim, Basel: Beltz Verlag.

Lübbert, D. (1999). Retrieved 01/10, 2013, from http://www.luebbert.net/uni/statist/zr/zr2.php

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau, 58*(2), 103-117.

Lumpe, A. T., & Staver, J. R. (1995). Peer Collaboration and Concept Development: Learning about Photosynthesis. *Journal of Research in Science Teaching, 32*(1), 71-98.

Markland, D., & Hardy, L. (1997). On the Factorial and Construct Validity of the Intrinsic Motivation Inventory: Conceptual and Operational Concerns. *Research Quarterly for Exercise and Sport, 68*(1), 20-31.

Martin, J. P. (1985). *Zum Aufbau didaktischer Teilkompetenzen beim Schüler: Fremdsprachenunterricht auf der lerntheoretischen Basis des Informationsverarbeitungsansatzes*. Tübingen: Gunter Narr Verlag.

Martin, J. P. (1998). Das Projekt "Lernen durch Lehren"–fachdidaktische Forschung im Spannungsfeld von Theorie und selbsterlebter Praxis. In M. Liedtke (Eds.), Gymnasium - Neue Formen des Unterrichts und der Erziehungpp. 151-166).

McAuley, E., Duncan, T., & Tammen, V. (1989). Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport, 60*(1), 48-58.

McAuley, E., Wraith, S., & Duncan, T. E. (1991). Self-Efficacy, Perceptions of Success, and Intrinsic Motivation for Exercise. *Journal of Applied Social Psychology, 21*, 139-155.

Miyake, N. (2008). Conceptual Change through Collaboration. In S. Vosniadou (Ed.), *International Handbook of Research on Conceptual Change* (pp. 453-478). New York: Poutledge.

Muckenfuß, H. (1995). *Physik im sinnstiftenden Kontext*. Berlin: Cornelsen.

Müller, F. H., Hanfstingl, B., & Andreitz, I. (2007). Skalen zur motivationalen Regulation beim Lernen von Schülerinnen und Schülern: Adaptierte und ergänzte Version des Academic Self-Regulation Questionnaire (SRQ-A) nach Ryan and Connell.

Wissenschaftliche Beiträge aus dem Institut für Unterrichts- und Schulentwicklung. Klagenfurt: Alpen-Adria-Universität.

Müller, M., Berger, R., & Hänze, M. (2013). Entwicklung von Trainings zur Verbesserung der Unterstützung im CAT. Unpublished Vortrag. GDCP Jahrestagung.

Müller, R. (2006). *Physik in interessanten Kontexten. Handreichung für die Unterrichtsentwicklung*. Kiel: IPN.

Neumann, S., & Hopf, M. (2012). Students' conceptions about 'radiation': Results from an explorative interview study of 9th grade students. *Journal of Science Education and Technology, 21*(6), 826-834.

OECD (Producer). (2010, 2012-09-05) PISA 2009 Ergebnisse: Zusammenfassung.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a Scientific Conception: Toward a Theory of Conceptual Change. *Science Education, 66*, 211-227.

Prediger, S., & Link, M. (2012). Fachdidaktische Entwicklungsforschung - ein lernprozessfokussierendes Forschungsprogramm mit Verschränkung fachdidaktischer Arbeitsbereiche. In H. Bayrhuber, U. Harms, B. Muszynski, B. Ralle, M. Rothgangel, L.-H. Schön, H. Vollmer & H.-G. Weigand (Eds.), *Formate Fachdidaktischer Forschung - Empirische Projekte-historische Analysen-theoretische Grundlagen* (Vol. Bd. 2). Münster: Waxmann.

Prenzel, M., Eitel, F., Holzbach, R., & Schönheinz, R.-J. (1993). Lernmotivation im studentischen Unterricht in der Chirurgie. *Zeitschrift für Pädagogische Psychologie, 7*(2/3), 125-137.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15.

Rechnungshof. (2013). Bericht des Rechnungshofes - Modellversuch Neue Mittelschule. Retrieved 05-31, 2014, from http://www.rechnungshof.gv.at/fileadmin/downloads/_jahre/2013/berichte/teilberichte/bund/Bund_2013_12/Bund_2013_12_1.pdf

Reusser, K. (2001). *Unterricht zwischen Wissensvermittlung und Lernen lernen*. Donauwörth: Auer.

Rheinberg, F. (2006). Motivationstraining und Motivation. In D. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie*. Weinheim: Beltz.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen 12 (Langversion, 2001). *Diagnostica, 2*, 57-66.

Riemeier, T. (2012). Moderater Konstruktivismus. In D. Krüger & H. Vogt (Eds.), *Theorien fachdidaktischer Forschung*.

Robinson, D. R., Schofield, J. W., & Steers-Wentzell, K. L. (2005). Peer and Cross-Age Tutoring in Math: Outcomes and Their Design Implications. *Educational Psychology Review, 17*(4), 327-362.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-

Assisted Learning Interventions With Elementary School Students: A Meta-Analytic Review. *Journal of Educational Psychology, 95*(2), 240-257.

Rost, D. (2007). *Interpretation und Bewertung pädagogisch-psychologischer Studien, 2. Auflage*. Weinheim, Basel: Beltz verlag.

Ryan, R. M. (1982). Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory. *Journal of Personality and Social Psychology, 43*(3), 450-461.

Ryan, R. M. (1995). Psychological Needs and the Facilitaion of Integrative Processes. *Journal of Personality, 36*(3), 397-427.

Ryan, R. M., & Connell, J. P. (1989). Perceived Locus of Causality and Internatization: Examining Reasons for Acting in Two Domains. *Journal of Personality and Social Psychology, 57*(5), 749-761.

Ryan, R. M., Connell, J. P., & Robert, W. (1990). Emotions in Non-directed Text Learning. *Learning and Individual Differences, 2*(1), 1-17.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology, 25*, 54-67.

Ryan, R. M., Koestner, R., & Deci, E. L. (1991). Varied forms of persistence: When free-choice behavior is not intrinsically motivated. *Motivation and Emotion, 15*, 185-205.

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology, 45*, 736-750.

Sachs, L. (2002). *Angewandte Statistik. Anwendung statistischer Methoden.* Berlin, Heidelberg, New York: Springer.

Schmirch, J. (2009). Eine Charakterisierung der Risikoschüler/innen. In C. Schreiner & U. Schwantner (Eds.), *PISA 2006. Österreichischer Expertenbericht zum Naturwissenschaftsschwerpunkt*. Graz: Leykam.

Schreiner, C., Schwantner, U. (Ed.). (2009). *PISA 2006. Österreichischer Expertenbericht zum Naturwissenschaftsschwerpunkt*. Graz: Leykam.

Schwantner, U., Toferer, B., & Schreiner, C. (Eds.). (2013). *PISA 2012: Internationaler Vergleich von Schülerleistungen: Erste Ergtebnisse Mathematik, Lesen, Naturwissenschaft*. Graz: Leykam.

Schwarzer, R. (1993). *Stess, Angst und Handlungsregulation.* Stuttgart: Kohlhammer.

Schwarzer, R., & Jerusalem, M. (1999a). Allgemeine Selbstwirksamkeitserwartung. Retrieved 07.11.2012, 2012

Schwarzer, R., & Jerusalem, M. (Eds.). (1999b). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen.* Berlin: Freie Universität Berlin.

Schwarzer, R., Mueller, J., & Greenglass, E. (1999). Assessment of perceived general self-efficacy on the Internet: Data collection in cyberspace. *Aniety, Stress and Coping, 12*, 145-161.

Shaffer, P. S., & McDermott, L. C. (1992). Research as a guide for curriculum development: An example from introductory electricity. Part II: Design of instructional strategies. *American Journal of Physics, 60*(11), 994-1013.

Shipstone, D. M. (1984). A study of children's understanding of electricity in simple DC circuits. *European Journal of Science Education, 6*(2), 185-198.

Statistik-Austria. (2013). Statistiken/Bildung, Kultur/Schülerinnen, Schüler. Retrieved 16/08, 2013, from http://www.statistik.at/web_de/statistiken/bildung_und_kultur/formales_bildungswesen/schulen_schulbesuch/index.html

Statistik-Austria. (2014). Schulen, Schulbesuch. Retrieved 01.06., 2014, from http://www.statistik.at/web_de/static/schulbesuch_der_5._schulstufe_200607_besuch_der_allgemeinbildenden_hoehere_035638.gif

Tabachnik, B. G., & Fidell, L. S. (2006). *Using multivariate statistics*. New York: Harper Collins College Publishers.

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education, 32*(3), 321-345.

Topping, K. J. (2005). Trends in Peer Learning. *Educational Psychology, 25*(6), 631-645.

Topping, K. J., Peter, C., Stephen, P., & Whale, M. (2004). Cross-age peer tutoring of science in the primary school: Influence on scientific language and thinking. *Educational Psychology, 24*(1), 57-75.

Treagust, D. F. (Ed.). (2007). *General instructional methods and strategies*. New York: Taylor and Francis.

Trinkl, C. (2012). *Lernprozesse zum Thema Schatten und Lichtausbreitung.* Universität Wien, Wien.

Tsigilis, N., & Theodosiou, A. (2003). Temporal Stability of the Intrinsic Motivation Inventory. *Perseptual and Motor Skills, 97*, 271-280.

Urban-Woldron, H., & Hopf, M. (2012). Testinstrument zum Verständnis in der Elektrizitätslehre. *Zeitschrift für Didaktik der Naturwissenschaften, Jg.18*.

Urhahne, D. (2008). Sieben Arten der Lernmotivation. *Psychologische Rundschau, 59*(3), 150-166.

Vallerand, R. J. (2000). Deci and Ryan's Self-Determination Theory: A View from the Hierarchical Model of Intrinsic and Extrinsic Motivation. *Psychological Inquiry, 11*(4), 312-318.

von Rhöneck, C. (1986). Vorstellungen vom elektrischen Stromkreis. *Naturwissenschaften im Unterricht Physik/Chemie, 34*(13), 10-14.

Vosniadou, S. (2013). Conceptual Change in Learning and Instruction. In S. Vosniadou (Ed.), *International Hanbook or Research on Conceptual Change. Second Edition*. New York: Routledge.

Vygotsky, L. S. (1978). *Mind in Society*. London: Harvard University Press.

Wayman, J. C. (2003). *Multiple Imputation for Missing Data: What is it and how can I use it?* Paper presented at the Annual Meeting of the American Educational Research

Association, Chicago.

Weiner, B., Frieze, I., Kukla, A., Reed, A., Rest, S., & Rosenbaum, L. M. (1971). *Perceiving the causes of success and failure*. Morristown: General Learning Press.

White, R., & Gunstone, R. (1992). *Probing Understanding*. London, New York: RoutledgeFalmer.

Whitehead, J. R., & Corbin, C. B. (1991). Youth Fitness Testing: The Effect of Percentile-Based Evaluative Feedback on Intrinsic Motivation. *Research Quarterly for Exercise and Sport, 62*(2), 225-231.

Widodo, A. (2004). *Constructivist Oriented Lessons. The Learning Environments and the Teaching Sequences* (Vol. 915). Frankfurt/Main: Peter Lang GmbH.

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

Wiesner, H. (1992). Schülervorstellungen und Lernschwierigkeiten mit dem Spiegelbild. *Naturwissenschaften im Unterricht - Physik, 3*(14), 16-18.

Wiesner, H. (2004a). Schülervorstellungen zur Elektrizitätslehre und Sachunterricht. In R. Müller, R. Wodzinski & M. Hopf (Eds.), *Schülervorstellungen in der Physik* (pp. 53-65). Köln: Aulis Verlag Deubner.

Wiesner, H. (2004b). Verbesserung des Lernerfolgs im Unterricht über Optik (I). In F. H. Müller, R. Wodzinski & M. Hopf (Eds.), *Schülervorstellungen in der Physik*. Köln: Aulis Verlag Deubner.

Wiesner, H. (2004c). Vorstellungen von Grundschülern über Schattenphänomene. In R. Müller, R. Wodzinski & M. Hopf (Eds.), *Schülervorstellungen in der Physik* (pp. 71-79). Köln: Aulis Verlag Deubner.

Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzskala intrinsischer Motivation (KIM). *Zeitschrift für Didaktik der Naturwissenschaften, 15*, 31-45.

Wilson, M. (2005). *Constructing Measures - An Item Response Modeling Approach*. New York: Taylor & Francis.

Zembylas, M. (2005). Three perspectives on linking the cognitive and the emotional in science learning: Conceptual change, socio-constructivism and poststructuralism. *Studies in Science Education, 41*, 91-115.

Zinn, B. (2008). *Physik lernen, um Physik zu lehren.* Universität Kassel, Kassel.

Zinn, B. (2009). Ergebnisse einer Pilotuntersuchung zur Unterrichtsmethode "Lernen durch Lehren". *Zeitschrift für Didaktik der Naturwissenschaften, Jg 15*, 325-329.

11. Abbildungsverzeichnis

Sofern die Abbildungen nicht selbst erstellt sind, habe ich mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

12. Anhänge A

12.1. Zusammenfassung (deutsch)

Die vorliegende Doktorarbeit dokumentiert die wesentlichsten Ergebnisse aus einer Studie über Cross-Age Peer Tutoring (CAPT) in Physik. Das bescheidene Abschneiden österreichischer Schüler/innen bei den OECD Studien war Grund genug, um auf die Suche nach Abhilfen und Maßnahmen zur Verbesserung des Unterrichts zu gehen. CAPT gilt in Schulkreisen als eine solche Maßnahme. In diesem Sinne wird in dieser Arbeit evaluiert, ob CAPT im Sinne konstruktivistischer Lerntheorien eine geeignete Lernumgebung darstellt um Physik zu vermitteln.

Die moderne Fachdidaktik versteht unter CAPT eine Lernform, bei der ältere Schüler/innen Jüngere beim Lernen unterstützen. Verglichen man mit den Anfängen, als Tutoren als Ersatzlehrer/innen verstanden wurden, zeigt sich hier bereits ein Wandel im Forschungsfokus, weg von den Tutees, hin zu den Tutoren.

Obwohl es einige Studien zur Wirksamkeit von CAPT in diversen Kontexten gibt, sind solche im naturwissenschaftlichen Kontext selten. Es ist a priori auch nicht klar, ob sich deren Ergebnisse ohne weiteres auf den Physikunterricht übertragen lassen, da naturwissenschaftliches Lernen, aufgrund des dabei notwendigen *Conceptual Change,* unter Umständen unterschiedlichen Rahmenbedingungen unterliegt als Lernen in anderen Gegenständen. Aus diesem Grund wurde der Forschungsfokus der hier präsentierten Arbeit auf Schüler/innen der Sekundarstufe 1 (10 bis 14-Jährige) und deren Wissenserwerb in ausgewählten physikalischen Bereichen der Elektrizitätslehre und der Optik gelegt. Darüber hinaus wurden Zusammenhänge zu einigen demografischen und motivationalen Parametern untersucht, genauso wie zur Rolle der Schüler/innen im Tutoringprozess und zur Klassenzugehörigkeit. Um motivationale Ergebnisse erfassen zu können wurde ein bestehender Fragebogen für die untersuchte Altersgruppe adaptiert. Es wurden an $N_E = 172$ (Elektrizitätslehre) und $N_O = 141$ (Optik) Schüler/innen klassenweise Tutorings, mehrheitlich in Dyaden, durchgeführt. Das Wissen der Schüler/innen wurde mit einem Praetest-Posttest-Follow-up Testdesign abgebildet. Parallel dazu wurden Pilotstudien zum Motivationsinstrument durchgeführt.

Die Ergebnisse zeigen, dass CAPT mit zufrieden stellenden Effektstärken zwischen 0,46 und 0,62 zu einer Verbesserung im Wissen der Schüler/innen beiträgt. Derartige Effekte können auch beobachtet werden, wenn der Anteil an Schüler/innen mit nicht-deutscher Muttersprache erheblich ist. Multiple lineare Regressionsmodelle zur Schätzung der Posttestscores in der Elektrizitätslehre zeigen, dass die einflussreichen Parameter die aktive Rolle im Tutoringprozess, die Muttersprache und der Praetestscore sind.

Hingegen zeigen Untersuchungen zu motivationalen Parametern lediglich schwache Korrelationen, obwohl das nach dem Studium der Literatur zu erwarten gewesen wäre. Analysen der Follow-up Tests weisen sowohl für die Elektrizitätslehre, als auch für die Optik auf eine zufrieden stellende Persistenz des Wissens hin.

Zusammenfassend konnte in der vorliegenden Dissertation nachgewiesen werden, dass CAPT zu Effekten bei allen beteiligten Lernenden führt.

12.2. Abstract (English)

This doctorial thesis presents the main results of a study conducted about cross-age peer tutoring (CAPT) in Physics. The study has been spurred by the search for remedies for the poor performance of Austrian students in science revealed in recent OECD studies, and for improvement in science teaching. It evaluates whether CAPT provides an appropriate constructivist-oriented learning environment for teaching Physics.

According to a review of literature about recent didactics, cross-age peer tutoring describes learning processes where older people from similar groups help younger ones to learn. Compared to the beginning of tutoring, when tutors were regarded as surrogate teachers, the research focus has shifted from tutees to tutors.

Though there is evidence from previous studies that CAPT works effectively in various contexts, studies about secondary level students within science contexts are rare. Results may not easily be transferred from other subjects because learning in science requires different learning environments in order to enhance the necessary conceptual change. Thus, the research focus of the present study has been put on on the academic achievement of secondary level 1 students (aged 10 to 14 years), concerning different contexts/selected fields of Physics (electricity and optics). Additionally, this academic achievement was linked to some demographic and motivational parameters, as well as

the students' role within the tutoring process and their affiliation to the different classes in school. In order to investigate motivational outcomes a well-known/existing inventory has been adapted for the respective age group.

N_E = 172 (electricity) and N_O = 141 (optics) students from grades 5 to 8 underwent class-wide cross-age peer tutoring processes, mostly in dyads. The students' academic outcomes were examined in a pretest-posttest-follow-up test design. Simultaneously, pilot studies on the motivational inventory were conducted.

The results show clearly that CAPT enhances the overall test results, compared to the pretests, with sufficient effect sizes between 0.46 and 0.62. This can also be observed even if there is a considerable percentage of students whose first language is not German. Multiple regression models estimating the posttest score in electricity reveal, that relevant parameters are the active role within the tutoring process, the first language, and the pretest score. Investigations on motivational parameters showed only weak correlations to the academic achievement, in contrast to what has been expected according to literature.

Analyses of the follow-up tests indicate that the achievement in electricity as well as in optics is satisfactorily persistent.

Summing up, the analyses of data have provided evidence that CAPT enhances knowledge for all participating learners.

12.3. Curriculum Vitae

geboren: 20. 6. 1968 in Wien

Ausbildung:

1986: Matura, GRG 15, Auf der Schmelz

1986 – 1992: Studium und Abschluss an der Universität Wien, Lehramt Physik, Mathematik

2005: ECHA Diplom (European Council of High Abilities), University of Nijmegen, Netherlands

seit 2010: PhD Studium am AECC Physik der Universität Wien

Tätigkeiten:

seit 1992: Lehrerin am pGRG 15, Friesgasse 4 in Wien

seit 2005: Mitarbeiterin im Schulentwicklungsteam

seit 2009: Landeskoordinatorin für Wien der Österreichischen Physikolympiade

2010 - 2013: Projektmitarbeiterin im Projekt Cross-Age Peer Tutoring in Physics 1

seit 2013: wissenschaftliche Mitarbeiterin am AECC Physik

seit 2011: Vorträge und Workshops an der PH Wien (Lehrer/innenfortbildung)

Forschungsinteressen:

Evaluationsforschung

Wirksamkeit von Cross Age Peer Tutoring in Physik

Motivationsstrategien

Fragebogenkonstruktion und statistische Analysen

Conceptual Change in der angewandten Unterrichtssituation

12.4. Publikationen

Cross-Age Peer Tutoring im Physikunterricht: Eine ungewöhnliche Unterrichtsmethode stellt sich vor, Korner, M. 2014 in: Plus Lucis. 1-2, S. 11-15

Cross-Age Peer Tutoring in Electricity: Comparing the Outcomes of Tutors and Tutees, Korner, M. & Hopf, M. 2014 *ESERA 2013, Nicosia, Cyprus: eProceedings.*

Cross-Age Peer Tutoring in Physics: Tutors, Tutees and Achievement in Electricity, Korner, M. & Hopf, M. 2014 in: International Journal of Science and Mathematics Education. 30 S.

Cross-Age-Peer Tutoring in Physik - Rolle und Lernerfolg, Korner, M., Urban-Woldron, H. & Hopf, M. 2013 *Inquiry-based Learning - Forschendes Lernen: GDCP, Jahrestagung in Hannover 2012.* Kiel: LIT Verlag, Band 33, (Gesellschaft für Didaktik der Chemie und Physik - Jahrestagung; Band 33)

Wellen unterrichten mit dem Internet, Hopf, M. & Korner, M. 2013 in : Praxis der Naturwissenschaft - Physik in der Schule.

Abschlussbericht zum Sparkling Science Projekt "Cross-Age Peer Tutoring in Physik", Korner, M., Urban-Woldron, H. & Hopf, M. 2012 Unknown publisher

Entwicklung eines Messinstrumentes zur Motivation, Korner, M., Urban-Woldron, H. & Hopf, M. 2012 *Konzepte fachdidaktischer Strukturierung für den Unterricht: GDCP, Jahrestagung in Oldenburg 2011.* LIT Verlag, Band 32, S. 98-100 3 S.

Peer Tutoring: Rollenverständnis und Lernprozesse, Korner, M., Urban-Woldron, H. & Hopf, M. 2012 *Konzepte fachdidaktischer Strukturierung für den Unterricht: GDCP, Jahrestagung in Oldenburg 2011.* Münster: LIT Verlag, Band 32, S. 646-648 3 S.

Cross Ager Peer Tutoring: Motivation und Wissenszuwachs, Abraham, D., Korner, M., Urban-Woldron, H. & Hopf, M. 2011, Veröffentlichung: Beitrag zu Konferenz, Poster

Peer Tutoring: Rollenverständnis und Lernprozesse, Himmer, B., Korner, M., Urban-Woldron, H. & Hopf, M. 2011, Veröffentlichung: Beitrag zu Konferenz, Poster

Radonbelastung in Wohnräumen. Kontextorientierte Aufgabe (16), Haagen-Schützenhöfer, C. & Korner, M. 2011 in : Praxis der Naturwissenschaft - Physik in der Schule. 60, 3, S. 36-38 3 S.

Zwischenberichtbericht zum Sparkling Science Projekt "Cross-Age Peer Tutoring in Physik", Korner, M. & Hopf, M. 2011

13. Anhänge B

13.1. Die IMI-Skalen in Originalform und in Übersetzung

Die Autoren Deci und Ryan (2003) bezeichnen diese Version ihres Fragebogens als „The Post-Experimental Intrinsic Motivation Inventory“. Auf der oben zitierten Homepage, die einen Überblick über sie SDT gibt und auf der Theorie, Artikel und Fragebögen zur Selbstbestimmungstheorie publiziert sind, sind auch die 45 unten angeführten Items zu finden, die sieben Subskalen bilden.

Die Autoren schlagen für die Testung eine 7-teilige Likertskala vor, beginnend mit 1 (= not at all true) bis zu 7 (= very true). Items, die mit (R) gekennzeichnet sind haben eine negative Polung, sind also mit „8 minus Wert“ ins Kalkül zu ziehen. Jene Items der *value/usefulness*-Skala, die unvollständig sind, sind so konstruiert, dass sie auf die betreffende Situation oder Tätigkeit angepasst werden können.

Unmittelbar nach jedem Originalitem ist auch die deutsche Übersetzung zu finden. Dabei handelt es sich um jene Version, die auf Basis von Übersetzung und Re-Übersetzung entstanden ist und bereits eine an Schüler/innen angepasste Sprache verwendet.

Interest/Enjoyment

Interesse/Vergnügen

int 1: *I enjoyed doing this activity very much*.

Ich habe diese Tätigkeit sehr gerne gemacht.

int 2: *This activity was fun to do.*

Diese Tätigkeit hat Spaß gemacht.

int 3: I thought this was a boring activity. (R)

Diese Tätigkeit war langweilig. (R)

int 4: *This activity did not hold my attention at all.* (R)

Ich schenkte dieser Tätigkeit überhaupt keine Aufmerksamkeit. (R)

int 5: *I would describe this activity as very interesting*.

Ich würde diese Tätigkeit als sehr interessant bezeichnen.

int 6: *I thought this activity was quite enjoyable*.

Diese Tätigkeit war recht vergnüglich.

int 7: *While I was doing this activity, I was thinking about how much I enjoyed it*.

Während ich diese Tätigkeit ausführte, dachte ich daran, wie sehr ich sie genossen habe.

Perceived Competence

(selbst) wahrgenommene Kompetenz

pco 1: *I think I am pretty good at this activity.*

Ich denke, ich bin ziemlich gut bei dieser Tätigkeit.

pco 2: *I think I did pretty well at this activity, compared to other students.*

Verglichen mit meinen Mitschüler/innen denke ich, dass ich bei dieser Tätigkeit recht gut war.

pco 3: After working at this activity for a while, I felt pretty competent.

Nach einer Weile fühlte ich mich recht fähig für diese Tätigkeit.

pco 4: *I am satisfied with my performance at this task.*

Ich bin mit meiner Darbietung bei diesen Aufgaben zufrieden.

pco 5: *I was pretty skilled at this activity*.

Ich bin recht geschickt bei dieser Tätigkeit.

pco 6: *This was an activity that I couldn't do very well.* (R)

Das war eine Tätigkeit, die ich nicht sehr gut bewältigen konnte. (R)

Effort/Importance

Einsatzbereitschaft / Wichtigkeit

eff 1: *I put a lot of effort into this.*

Ich habe mich sehr eingesetzt bei dieser Tätigkeit.

eff 2: *I didn't try very hard to do well at this activity.* (R)

Ich habe mich nicht sehr angestrengt diese Tätigkeit gut zu machen. (R)

eff 3: *I tried very hard on this activity.*

Ich habe mich sehr angestrengt bei dieser Tätigkeit.

eff 4: *It was important to me to do well at this task.*

Es war wichtig für mich, diese Aufgabe gut zu bewältigen.

eff 5: *I didn't put much energy into this.* (R)

Ich habe nicht viel Energie hineingesteckt. (R)

Pressure/Tension

Druck/Anspannung

I did not feel nervous at all while doing this. (R)

Ich war überhaupt nicht nervös während ich das tat. (R)

I felt very tense while doing this activity.

Ich war sehr angespannt bei dieser Tätigkeit

I was very relaxed in doing these. (R)

Ich war sehr entspannt dabei. (R)

I was anxious while working on this task.

Ich hatte bei diesen Aufgaben Angst.

I felt pressured while doing these.

Ich fühlte mich bei dieser Tätigkeit unter Druck.

Perceived Choice

wahrgenommene Wahlfreiheit

pch 1: *I believe I had some choice about doing this activity.*

Ich glaube, dass ich bei dieser Tätigkeit Wahlfreiheit hatte.

pch 2: *I felt like it was not my own choice to do this task.* (R)

Ich fühlte, dass ich nicht entscheiden konnte, ob ich das mache oder nicht. (R)

pch 3: *I didn't really have a choice about doing this task.* (R)

Ich hatte keine andere Wahl als diese Aufgabe zu erledigen. (R)

pch 4: *I felt like I had to do this.* (R)

Mir kam vor, dass ich das tun musste. (R)

pch 5: *I did this activity because I had no choice.* (R)

Ich machte diese Tätigkeit, weil ich keine andere Wahl hatte. (R)

pch 6: *I did this activity because I wanted to.*

Ich machte diese Tätigkeit, weil ich es wollte.

pch 7: *I did this activity because I had to.* (R)

Ich machte diese Tätigkeit, weil ich musste. (R)

Value/Usefulness

Nützlichkeit/Wert

val 1: *I believe this activity could be of some value to me.*

Ich glaube, dass mir diese Tätigkeit etwas bringen könnte.

val 2: *I think that doing this activity is useful for* _____

Ich denke, dass diese Tätigkeit brauchbar ist um _____

val 3: *I think this is important to do because it can* _____

Ich denke das ist wichtig weil ich _____ kann.

val 4: *I would be willing to do this again because it has some value to me.*

Ich bin bereit das wieder zu machen, weil es für mich wertvoll ist.

val 5: *I think doing this activity could help me to* _____

Ich denke, dass mir diese Tätigkeit helfen könnte um _____

val 6: *I believe doing this activity could be beneficial to me.*

Ich denke es könnte vorteilhaft für mich das zu tun.

val 7: *I think this is an important activity.*

Ich denke das ist eine wichtige Tätigkeit.

Relatedness

Eingebundenheit

rel 1: *I felt really distant to this person.* (R)

Ich fühlte mich der Person richtig fern. (R)

rel 2: *I really doubt that this person and I would ever be friends.* (R)

Ich bezweifele, dass diese Person und ich jemals Freunde werden könnten. (R)

rel 3: *I felt like I could really trust this person.*

Ich fühlte, dass ich dieser Person wirklich vertrauen konnte.

rel 4: *I'd like a chance to interact with this person more often.*

Ich würde gerne eine Gelegenheit haben mit dieser Person öfter zusammenzuarbeiten.

rel 5: *I'd really prefer not to interact with this person in the future.* (R)

Ich würde mit dieser Person lieber nicht mehr zusammenarbeiten. (R)

rel 6: *I don't feel like I could really trust this person.* (R)

Ich habe nicht das Gefühl, dass ich dieser Person wirklich vertrauen kann. (R)

rel 7: *It is likely that this person and I could become friends if we interacted a lot.*

Es ist möglich, dass diese Person und ich Freunde werden, wenn wir viel zusammenarbeiten.

rel 8: *I feel close to this person.*

Ich fühle mich dieser Person nahe.

13.2. Online-Version des Motivationsfragebogens für die erste Pilotierung

Aktivitäten im Unterricht

Liebe Schülerin, lieber Schüler!

Wir bitten Dich diesen Fragebogen für uns zu testen. Er hat sehr viele ähnliche Fragen, das ist Absicht. Wir wollen mit Deiner Hilfe herausfinden, welche davon am besten sind. Denk bei "dieser Tätigkeit" an eine Gruppenarbeit, einen Theaterworkshop, ein Arbeitsblatt, eine Exkursion oder an eine beliebige Stunde. Keine Sorge, Deine Antworten sind ganz anonym. Bitte kreuze bei JEDER Frage EINE Möglichkeit an.

1. Gib bitte die Klasse und die Schultype an (zB 3HS für 3. Klasse Hauptschule)

_____

2. Ich habe diese Tätigkeit sehr gerne gemacht.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

3. Ich denke ich bin ziemlich gut bei dieser Tätigkeit.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

4. Ich habe mich bei dieser Tätigkeit sehr eingesetzt.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

5. Ich glaube, dass ich bei dieser Tätigkeit Wahlfreiheit hatte.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

6. Ich glaube, dass ich mir diese Tätigkeit etwas bringen könnte.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

7. Ich denke, dass diese Tätigkeit brauchbar ist um in Physik/Biologie/Geschichte/... besser zu verstehen.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

8. Ich machte diese Tätigkeit, weil ich musste.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

9. Es war wichtig für mich, diese Aufgabe gut zu bewältigen.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

10. Ich bin recht geschickt bei dieser Tätigkeit.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

11. Diese Tätigkeit war recht vergnüglich.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

12. Ich würde diese Tätigkeit als interessant bezeichnen.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

13. Verglichen mit meinen MitschülerInnen denke ich, dass ich bei dieser Tätigkeit recht gut war.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

14. Ich habe mich sehr angestrengt bei dieser Tätigkeit.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

15. Ich hatte keine andere Wahl als diese Aufgabe zu erledigen.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

16. Ich denke das ist wichtig, weil ich dadurch das Thema besser verstehen kann.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

17. Ich denke das ist eine wichtige Tätigkeit.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

18. Ich machte diese Tätigkeit, weil ich es wollte.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

19. Das war eine Tätigkeit, die ich nicht sehr gut bewältigen konnte.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

20. Während ich diese Tätigkeit ausführte dachte ich daran, wie sehr ich sie genossen habe.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

21. Diese Tätigkeit hat Spaß gemacht.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

22. Nach einer Weile fühlte ich mich recht fähig für diese Tätigkeit.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

23. Ich habe nicht viel Energie hineingesteckt.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

24. Ich machte diese Tätigkeit, weil ich keine andere Wahl hatte.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

25. Ich denke, dass mir diese Tätigkeit helfen könnte um in Physik/Biologie/Geschichte/... besser zu werden.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

26. Mir kam vor, dass ich das tun musste.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

27. Ich habe mich nicht sehr angestrengt diese Tätigkeit gut zu machen.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

28. Ich bin mit meiner Darbietung bei diesen Aufgaben zufrieden.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

29. Diese Tätigkeit war langweilig.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

30. Ich fühlte, dass ich nicht entscheiden konnte, ob ich das mache oder nicht.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

31. Ich bin bereit das wieder zu machen, weil es für mich wertvoll ist.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

32. Ich schenkte dieser Tätigkeit überhaupt keine Aufmerksamkeit.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

33. Ich denke, es könnte für mich vorteilhaft sein das zu tun.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

Bitte denke für die folgenden Fragen an eine Person (vielleicht aus deiner Klasse), mit der du in letzter Zeit zusammen gearbeitet hast (zB in einer Gruppenarbeit).

34. Ich fühlte mich der Person richtig fern.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

35. Ich bezweifle, dass diese Person und ich jemals Freunde werden könnten.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

36. Ich fühlte, dass ich dieser Person wirklich vertrauen konnte.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

37. Ich würde gerne eine Gelegenheit haben mit dieser Person öfter zusammenzuarbeiten.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

38. Ich würde mit dieser Person lieber nicht mehr zusammenarbeiten.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

39. Ich habe nicht das Gefühl, dass ich dieser Person wirklich vertrauen kann.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

40. Es ist möglich, dass diese Person und ich Freunde werden, wenn wir viel zusammenarbeiten.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

41. Ich fühle mich dieser Person nahe.
a☐ stimmt völlig b☐ stimmt eher c☐ stimmt teilweise d☐ stimmt eher nicht e☐ stimmt gar nicht

Danke vielmals für die Beantwortung aller Fragen!

13.3. Motivationsfragebogen kurz nach der ersten Pilotierung

Der hier angegebene Fragebogen entstand aus der ersten Pilotierung des unter 13.2 angegebenen Fragebogens, nachdem redundante bzw. für Schüler/innen unverständliche Items gelöscht oder umformuliert wurden (vgl. 6.4). Er wurde im ersten Studienjahr so, wie er hier abgedruckt ist, an Schülerinnen und Schüler verteilt.

Liebe Schülerin, lieber Schüler!

Die folgenden Fragen beziehen sich auf das eben gemachte ***Tutoring****. Denke an die Schülerin / den Schüler, mit dem du gerade gearbeitet hast. Bitte versuche möglichst genau zu antworten.*
Alle Antworten werden anonym behandelt.

		stimmt völlig	**stimmt eher**	**stimmt teilweise**	**stimmt eher nicht**	**stimmt gar nicht**
1	Diese Tätigkeit hat Spaß gemacht.	o	o	o	o	o
2	Ich konnte den anderen Schülerinnen und Schülern wirklich vertrauen.	o	o	o	o	o
3	Ich würde gerne Gelegenheit haben mit ihr / ihm öfter zusammenzuarbeiten.	o	o	o	o	o
4	Ich konnte auswählen, wie ich es ihm / ihr erkläre.	o	o	o	o	o
5	Ich habe mich angestrengt diese Tätigkeit gut zu machen.	o	o	o	o	o
6	Ich fand diese Tätigkeit sehr interessant.	o	o	o	o	o
7	Ich konnte mitentscheiden, welche Aufgaben ich beim Tutoring mache.	o	o	o	o	o
8	Ich habe da viel Energie hineingesteckt.	o	o	o	o	o
9	Ich dachte beim Tutoring daran, wie gerne ich anderen etwas erkläre.	o	o	o	o	o
10	Es ist möglich, dass das andere Kind und ich Freunde werden, wenn wir viel zusammenarbeiten.	o	o	o	o	o
11	Ich machte hier mit, weil ich keine andere Wahl hatte.	o	o	o	o	o
12	Ich habe mich bei dieser Tätigkeit bemüht.	o	o	o	o	o
13	Ich war beim Erklären recht gut, wenn ich mich mit meinen Mitschülern vergleiche.	o	o	o	o	o
14	Ich glaube, dass ich bei dieser Tätigkeit eigene Entscheidungen treffen konnte.	o	o	o	o	o
15	Ich konnte spüren, wie es der anderen	o	o	o	o	o

	Schülerin / dem anderen Schüler ging.					
16	Ich habe diese Tätigkeit sehr gerne gemacht.	o	o	o	o	o
17	Ich habe mich beim Tutoring sehr angestrengt.	o	o	o	o	o
18	Ich bin mit meiner Leistung beim Erklären zufrieden.	o	o	o	o	o
19	Ich bin bereit das wieder zu machen.	o	o	o	o	o
20	Ich machte diese Tätigkeit, weil ich es selbst wollte.	o	o	o	o	o
21	Ich war recht geschickt bei dieser Tätigkeit.	o	o	o	o	o
22	Es war wichtig für mich diese Aufgabe gut zu bewältigen.	o	o	o	o	o
23	Ich machte das Tutoring, weil ich musste.	o	o	o	o	o

13.4. Motivationsfragebogen lang für die zweite Pilotierung

Der hier angegebene Fragebogen wurde im zweiten Studienjahr so, wie er hier abgedruckt ist, an Schülerinnen und Schüler verteilt. Dabei wurden den alten Effort-Items aus dem IMI (eff 1 bis eff 5) neu konstruierte Items hinzugefügt (eff 6 bis eff 17). Details der Itemkonstruktion sind Kapitel 6.5 zu entnehmen.

Einen Überblick über die Itemnummern des Motivationsfragebogens kurz (Anhang 13.3) im Vergleich zum Motivationsfragebogen lang (Anhang 13.4) gibt die folgende Tabelle 13.1:

FB Motivation kurz	FB Motivation lang	Item/Subskala	FB Motivation kurz	FB Motivation lang	Item/Subskala
mot 1	mot 1	int 2		mot 19	eff 12
mot 2	mot 2	rel 3	mot 13	mot 20	pco 2
	mot 3	eff 6		mot 21	eff 13
mot 3	mot 4	rel 4	mot 14	mot 22	pch 1
	mot 5	eff 7	mot 15	mot 23	rel 8
mot 4	mot 6	pch 3	mot 16	mot 24	int 1
mot 5	mot 7	eff 2		mot 25	eff 14
	mot 8	eff 8	mot 17	mot 26	eff 3
mot 6	mot 9	int 5	mot 18	mot 27	pco 4
mot 7	mot 10	pch 4		mot 28	eff 15
	mot 11	eff 9	mot 19	mot 29	val 4
mot 8	mot 12	eff 5	mot 20	mot 30	pch 6
mot 9	mot 13	int 7		mot 31	eff 16
	mot 14	eff 10	mot 21	mot 32	pco 5
mot 10	mot 15	rel 7	mot 22	mot 33	eff 4
mot 11	mot 16	pch 5 (R)		mot 34	eff17
	mot 17	eff 11	mot 23	mot 35	pch 7 (R)
mot 12	mot 18	eff 1			

Tabelle 13.1: Überblick über die Nummerierungen in beiden Fragebogenversionen zur Motivation und die Zuordnung der Items bezüglich der Subskalen des IMI.

Liebe Schülerin, lieber Schüler!

Die folgenden Fragen beziehen sich auf eine eben gemachte eben gemachte Arbeit, z.B. eine Gruppenarbeit. Bitte versuche möglichst genau zu antworten.
Alle Antworten werden anonym behandelt.

		stimmt völlig	**stimmt eher**	**stimmt teilweise**	**stimmt eher nicht**	**stimmt gar nicht**
1	Diese Tätigkeit hat Spaß gemacht.	o	o	o	o	o
2	Ich konnte ihr / ihm wirklich vertrauen.	o	o	o	o	o
3	Ich übernehme auch schwierige Aufgaben.	o	o	o	o	o
4	Ich würde gerne Gelegenheit haben mit ihr / ihm öfter zusammenzuarbeiten.	o	o	o	o	o
5	Ich habe nachgefragt, bis der andere/die andere wirklich alles verstanden hat.	o	o	o	o	o
6	Ich konnte auswählen, wie ich es ihm / ihr erkläre.	o	o	o	o	o
7	Ich habe mich angestrengt diese Tätigkeit gut zu machen.	o	o	o	o	o
8	Ich habe bis zum Schluss der Arbeitszeit an dieser Aufgabe gearbeitet.	o	o	o	o	o
9	Ich fand diese Tätigkeit sehr interessant.	o	o	o	o	o
10	Ich konnte mitentscheiden, welche Aufgaben ich hier mache.	o	o	o	o	o
11	Ich habe mich sofort auf meine Aufgaben gestürzt.	o	o	o	o	o
12	Ich habe da viel Energie hineingesteckt.	o	o	o	o	o
13	Ich dachte beim Tutoring daran, wie gerne ich anderen etwas erkläre.	o	o	o	o	o
14	Ich habe bei dieser Aufgabe viel gearbeitet und bin nicht nur herumgesessen.	o	o	o	o	o
15	Es ist möglich, dass das andere Kind und ich Freunde werden, wenn wir viel zusammenarbeiten.	o	o	o	o	o
16	Ich machte hier mit, weil ich keine andere Wahl hatte.	o	o	o	o	o
17	Ich habe weiter gearbeitet, bis ich mit dem Ergebnis zufrieden war.	o	o	o	o	o

18	Ich habe mich bei dieser Tätigkeit bemüht.	o	o	o	o	o
19	Ich habe bei dieser Aufgabe kaum an etwas anderes gedacht.	o	o	o	o	o
20	Ich war beim Erklären recht gut, wenn ich mich mit meinen Mitschülern vergleiche.	o	o	o	o	o
21	Ich habe sogar auf etwas Freizeit (Pause) verzichtet um bei dieser Aufgabe gut zu sein.	o	o	o	o	o
22	Ich glaube, dass ich bei dieser Tätigkeit eigene Entscheidungen treffen konnte.	o	o	o	o	o
23	Ich konnte spüren, wie es der anderen Schülerin / dem anderen Schüler ging.	o	o	o	o	o
24	Ich habe diese Tätigkeit sehr gerne gemacht.	o	o	o	o	o
25	Ich habe versucht offenen Fragen auf den Grund zu gehen.	o	o	o	o	o
26	Ich habe mich hier sehr angestrengt um gut zu sein.	o	o	o	o	o
27	Ich bin mit meiner Leistung beim Erklären zufrieden.	o	o	o	o	o
28	Ich habe bei dieser Aufgabe kaum Pausen gemacht.	o	o	o	o	o
29	Ich bin bereit das wieder zu machen.	o	o	o	o	o
30	Ich machte diese Tätigkeit, weil ich es selbst wollte.	o	o	o	o	o
31	Ich habe mir nie gedacht: „Wann kann ich endlich aufhören?“	o	o	o	o	o
32	Ich war recht geschickt bei dieser Tätigkeit.	o	o	o	o	o
33	Es war wichtig für mich diese Aufgabe gut zu bewältigen.	o	o	o	o	o
34	Am liebsten hätte ich noch nicht aufgehört daran weiter zu arbeiten.	o	o	o	o	o
35	Ich machte diese Arbeit, weil ich musste.	o	o	o	o	o

13.5. Materialien für das Mentoring

Beispielhafte Auswahl zu Aufgabenstellungen für das Mentoring zum Thema Elektrizitätslehre.

Alle Fragestellungen sind nach folgendem Prinzip zu bearbeiten:
Vorhersage – experimentelle Überprüfung – Erklärung
Suche bitte Aufgaben aus, die deiner Meinung nach für die Volksschule geeignet sind. ***Beschränke*** *dich dabei auf* ***2 Kapitel!***

Problemstellungen für Kindergarten und Volksschule

Stromkreis

1) Flachbatterie und Lämpchen:
   Frage: Wann leuchtet Lämpchen? → Verwende die Abbildungen auf den kleinen Zetteln dafür.
2) Geschicklichkeitsspiel: 1 Lämpchen, 1 Draht, 1 Batterie. Bringe das Lämpchen zum Leuchten!
3) Verwende Kabel und Fassung und teste verschiedene Fragestellungen:
   Was passiert, wenn man eine andere Kabelfarbe / Kabellänge nimmt? ... Knoten / Schlingen macht?
   ➔ Diskutiere die Frage: Woher kommt der Name „Stromkreis"?
4) Heißer Draht (siehe Abbildung 1, Modell ist im Kindergartenvorhanden) → Zeichne / fahre mit dem Finger den Stromkreis nach.
5) Was passiert, wenn man in einem Stromkreis ein anderes Lämpchen verwendet / Anschlüsse beim Lämpchen vertauscht / Anschlüsse bei der Batterie vertauscht?
6) Untersuche die Anschlüsse bei der Batterie!
   Was stellst du fest?
7) Verwende einen Motor statt eines Lämpchens.
   Wie dreht er sich?
   Wie dreht er sich, wenn man die Anschlüsse am Motor vertauscht?
   Wie dreht er sich, wenn man die Anschlüsse an der Batterie vertauscht?
   Warum ist das so?
8) Wie sieht eine Glühbirne im Inneren aus? Untersuche die Glühbirne mit der Lupe und entwickle einen Bauplan! Verwende dazu die Abbildung 2.
   Baue eine Schaltung mit einer Glühbirne: Zeige und zeichne den Weg des Stromes.
9) Frage: Wo ist der Stromkreis bei Elektrogeräten? →??

Schalter

1) Schalter in den Stromkreis einbauen (Büroklammer)
   Macht es einen Unterschied, ob der Schalter vorher oder nachher eingebaut ist?
2) Den Schalter kann man auf 2 verschiedene Arten einbauen (bzw. 2 Schalter verwenden):
   Frage: Kannst du das Lämpchen ein- und ausschalten?

Leiter / Nichtleiter

1) Welche Materialien leiten Strom? →Teste verschiedene Materialien.

Abbildung 1: Heißer Draht
Quelle: http://www.gymnasium-wk.de/sites/default/files/Hei%C3%9Fer%20Draht%20Ulrike%20Gei%2013%20red%20400.jpg, 10.06.2014

Abbildung 2: Quelle: http://www.cvauf.de/download_main.htm, 24.10.2010

Beispielhafte Auswahl zu Aufgabenstellungen für das Mentoring zum Thema Optik (Schatten).

4) Schau auf die Zeichnung. Was hat der Zeichner nicht gewusst? Zeichne den Schatten richtig ein!

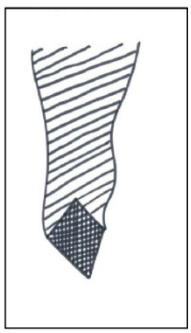


Abbildung 3: selbst nach: http://ne.lo-net2.de/selbstlernmaterial/p/s1o/ls/ls_au.pdf

5) Betrachte die Zeichnung! Welches Auge kann die Kerze sehen? – Kreise es ein!

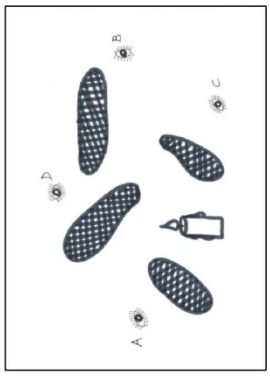


Abbildung 4: selbst nach: http://ne.lo-net2.de/selbstlernmaterial/p/s1o/ls/ls_au.pdf

Beispielhafte Auswahl zu Aufgabenstellungen für das Mentoring zum Thema Optik (Spiegel).

9) Wir wollen einen Schatten erzeugen... KDG, VS
Kinder versuchen einen Schatten zu erzeugen.
Was brauchst du dafür?
Lichtquelle (Lampe,...)
Gegenstand (=Ding,...)
ebene (Projektions-) Fläche (=Wand,...)
Fragen:
- Was ist der Schatten?
- Wie entsteht Schatten?
- In welche Richtung fällt der Schatten? → Supra Material „Fragen und Antworten"

Weitere Möglichkeiten:
- "Schattenfangen": Berühren des Schattens des Mitspielers mit dem Fuß oder mit dem Schatten der Hand.
- "Schatten verstecken": Kannst du deinen Schatten in einem anderen Schatten verstecken.
- "Schatten zeichnen": Aufzeichnen des Schattenumrisses mit Kreide oder auf Papier.
- Kannst du dich bewegen, ohne dass sich dein Schatten bewegt?
- Kannst du deinem Schatten auf den Kopf springen?
- Partnerarbeit: Unsere Schatten geben sich die Hand.

10) Große und kleine Schatten KDG, VS, SEK 1
Schatten mit Stofftieren (klein!)

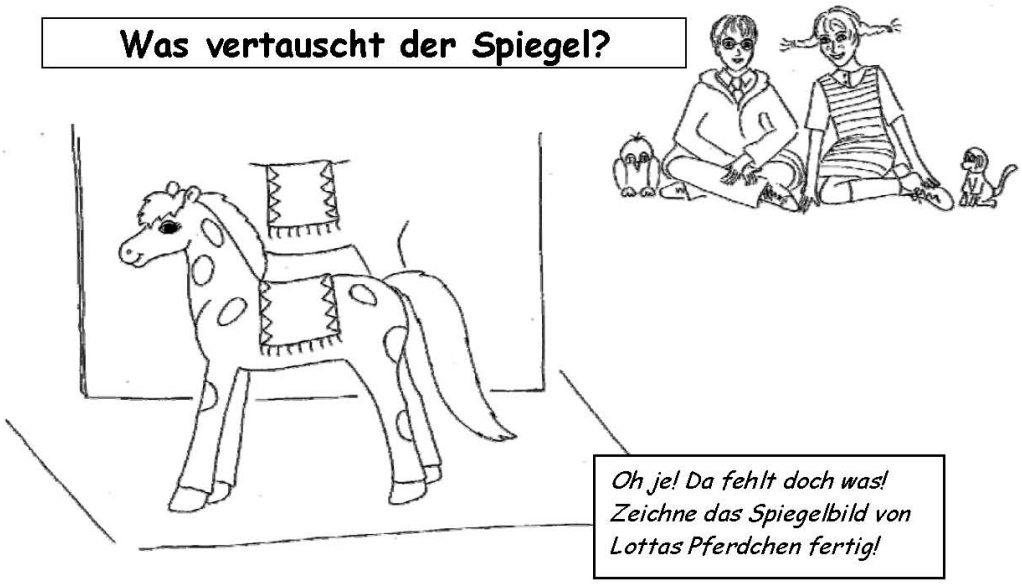


Abbildung 13.1: Eine Aufgabe um die Vorstellung zu schulen, dass der Spiegel Vorder- und Rückseite eines Körpers vertauscht.
Quelle: http://www.supra-lernplattform.de/index.php/lernfeld-natur-und-technik/spiegel, 12.2.2014

13.6. Materialien für das Tutoring

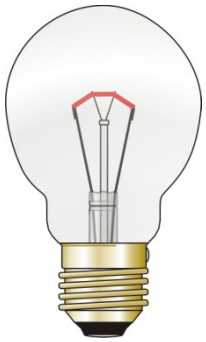
Beispielmaterialien für die Verwendung in Form von Aufgabenkärtchen im Rahmen von CAPT. Abbildung 12. und Abbildung 12. wurden im Bereich der Elektrizitätslehre eingesetzt, Abbildung 12. und Abbildung 12.2 im Bereich der Optik.

Teste deine Geschicklichkeit!

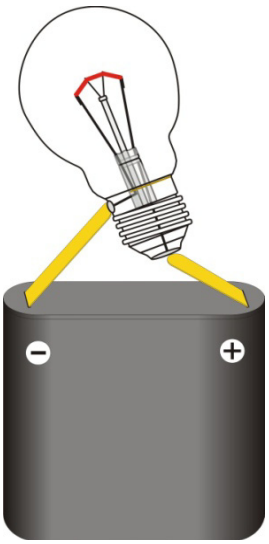
Kannst du das Lämpchen zum Leuchten bringen?

Abbildung 13.2: Eine Aufgabe, die die Vorstellung des geschlossenen Stromkreises schulen soll.

Weshalb leuchtet die Lampe?

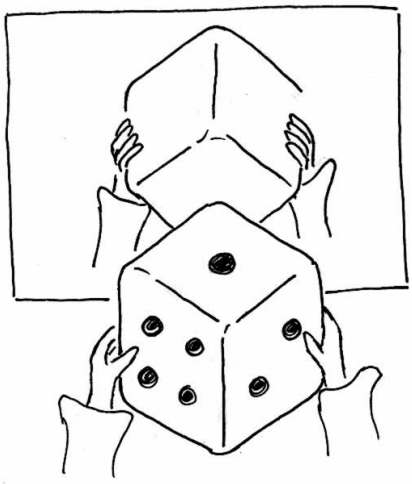


Wie sehen die Anschlüsse im Inneren der Lampe aus?



Wie fließt der Strom durch Batterie und Glühlampe? Zeichne seinen Weg mit Farbe ein!

Abbildung 13.3: Zur Vorstellung des geschlossenen Stromkreises



Zeichne die fehlenden Punkte in das Spiegelbild des Würfels ein!

Hast du richtig eingezeichnet?
Überprüfe dein Ergebnis mit einem Versuch!

Abbildung 13.4: Eine Aufgabe um die Vorstellung zu schulen, dass der Spiegel Vorder- und Rückseite eines Körpers vertauscht. Diese Tutoring-Aufgabe korrespondiert mit der Mentoring-Aufgabe aus Abbildung 13.1: Eine Aufgabe um die Vorstellung zu schulen, dass der Spiegel Vorder- und Rückseite eines Körpers vertauscht.
Quelle: http://www.supra-lernplattform.de/index.php/lernfeld-natur-und-technik/spiegel, 12.2.2014

Lösung V6:

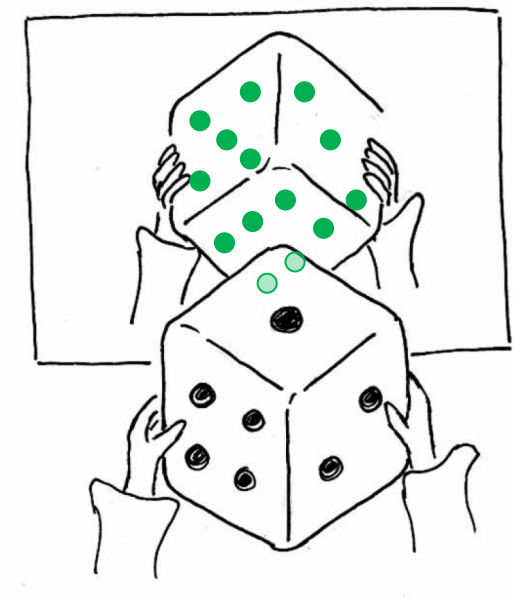
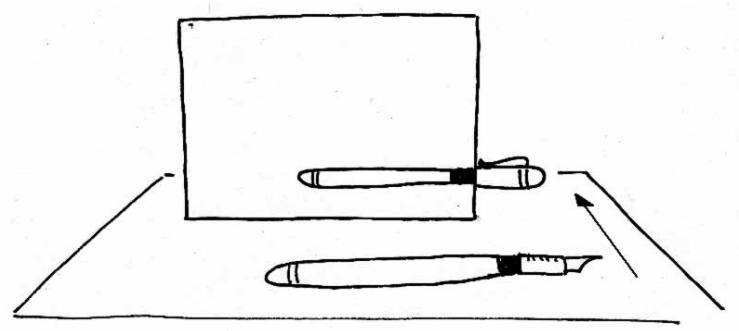


Abbildung 13.5: Lösung zur obigen Aufgabe. Diese Abbildung stellt die Rückseite des Hilfskärtchens dar.

Wo sehen wir das Spiegelbild?



Ziehe die Kappe von einem Stift ab.
Lege den Stift so vor den Spiegel wie in der Abbildung.
Nimm die Kappe des Stiftes und verschiebe sie so weit nach hinten, bis die Kappe und das Spiegelbild wie ein vollständiger Stift aussehen.

Miss nun den Abstand vom Stift zum Spiegel und vom Spiegel zur Kappe. **Was stellst du fest?**

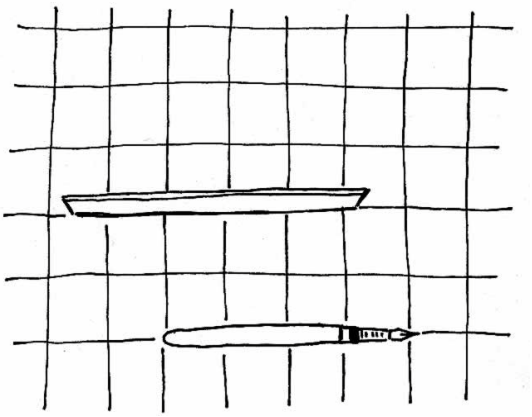


Abbildung 13.2: Zum Ort des Spiegelbildes hinter dem Spiegel.
Quelle: http://www.supra-lernplattform.de/index.php/lernfeld-natur-und-technik/spiegel, 12.2.2014

14. Dankesworte

Ich danke.

An aller erster Stelle meinem Betreuer Martin Hopf, der mich nicht nur ermutigt hat, eine derartige Arbeit in Angriff zu nehmen, sondern auch in allen möglichen Belangen bestmöglich unterstützt hat.

Meinen Kolleg/innen Ingrid Krumphals, Susanne Neumann und Herrn Doktor, Claudia Haagen-Schützenhöfer und Dominik Ertl, für viele wunderbare Diskussionen über MLRs, Formulierungen, im Deutschen, wie im Englischen, Itemfindung, NOS und für alle Unterstützung bei Vorträgen und bei der Datensammlung. Kurz: für gemeinsames Arbeiten und Probleme bewältigen.

Nochmals meinen Bürokolleg/innen, dafür, dass wir einander das Büro und die Arbeit dort immer wieder versüßt haben, und das nicht nur mit Aperolgummibären.

Brigitte Hauser für ihre Hilfe bei den englischsprachigen Formulierungen.

Meinen Kindern, die immer wieder ein Ich-bin-kurz-im-Büro und dann lange weg ertragen durften.

Meinem Bruder, für Hilfestellungen in der Statistik und die Eröffnung der Möglichkeit, Äpfel mit Birnen zu vergleichen.

Meinem Vater, der sich immer wieder über meine Arbeit erzählen ließ.

Meiner Mutter, nicht nur für hunderte hinzugefügte Beistriche beim Korrekturlesen, sondern auch für alles Interesse an meiner Arbeit.

Mein spezieller Dank gilt dem stanna-free-wlan.

Die vorliegende Studie wurde vom BMBF (vormals BMUKK) im Rahmen eines Sparkling Science Projektes unterstützt.