2 Empirische Untersuchungen

2.1 Theoretische Vorüberlegungen zum Lerngegenstand

2.1.1 Gründe für die Wahl des Lerngegenstands

Die im zweiten Teil dieser Arbeit vorgestellten empirischen Studien (Pilot- und Hauptstudie) fanden im Bereich der Strahlenoptik statt. Mehrere Gründe sprachen für die Entscheidung, Strahlenoptik als Lerngegenstand zu wählen.

1. Viele Studien im Bereich des Umgangs mit multiplen Repräsentationen im Hinblick auf die Förderung des konzeptuellen Verständnisses fanden im Bereich der Mechanik / Teilchenmodelle (Plötzner & Spada, 1998; Waldrip et al., 2010; Hubber et al., 2010) bzw. der Kinematik (Wilhelm, 2005) statt. Im Vergleich hierzu ist das Gebiet der Strahlenoptik weniger stark erforscht. Die in dieser Arbeit vorgestellten Studien im Bereich der Strahlenoptik wurden mit Oberstufenschülern (vgl. Mortimer & Buty, 2009) oder College-Studenten (vgl. Goldberg & McDermott, 1987) durchgeführt.
2. Das Gebiet der Strahlenoptik weist für die Untersuchung den „technischen" Vorteil auf, dass die Unterschiede in den Vorkenntnissen der Schüler kaum oder in geringem Maße durch den vorherigen Unterricht bedingt sind. Zum einen liegt dies daran, dass das Themengebiet der Strahlenoptik entsprechend dem Lehrplan für das Land Rheinland-Pfalz (vgl. Ministerium für Bildung, Wissenschaft, Jugend und Kultur. Rheinland-Pfalz. Lehrplan-Entwürfe Lernbereich Naturwissenschaften Biologie Physik Chemie S. 175 f., 190) das erste Gebiet ist, in dem die Schüler im Fach Physik an Gymnasien und Realschulen unterrichtet werden. Selbst wenn ein anderes Gebiet zeitlich vorgezogen wurde, fällt dies wenig ins Gewicht, da „das Gebiet der Optik […] wenig weitergehende Bezüge zu den nachfolgenden Lerninhalten aufweist" (zit. n. ebd., S. 190).
3. Die Strahlenoptik stellt entsprechend ein relativ in sich geschlossenes Gebiet dar, in dem die wesentlichen zentralen Grundstrukturen der Physik enthalten sind (zit. n. ebd.):

R. Hettmannsperger, *Lernen mit multiplen Repräsentationen aus Experimenten,*
DOI 10.1007/978-3-658-07436-4_2, © Springer Fachmedien Wiesbaden 2015

> „Ausgehend von der Beobachtung von Phänomenen der Alltagswelt sollen die Kinder einen ersten Einblick in die Untersuchung optischer Erscheinungen gewinnen. […] Gerade das Gebiet der Optik, das wenig weitergehende Bezüge zu den nachfolgenden Lerninhalten aufweist, […] ist gut geeignet, die Kinder behutsam in die Arbeitsweisen der Physik einzuführen. Elemente der physikalischen Fachsprache und Begriffswelt werden eingeübt. Am Beispiel des Lichts wird zum ersten Mal eine sinnvolle Modellbildung aufgezeigt. Das Experiment steht im Mittelpunkt des Unterrichts. Wo es nur angeht, sollen die Kinder selbst tätig werden“ (S. 190).

4. Die Strahlenoptik enthält somit die grundlegenden wissenschaftstheoretischen Aspekte und Methoden der Physik als wissenschaftliche Disziplin: Experimentieren, Beobachten und Modellbildung. Die Bildentstehung durch die Sammellinse wird hierbei explizit im Lehrplan für Realschulen und Gymnasien als ein zentraler Themenbereich genannt. Die folgenden Lerninhalte sind gemäß Lehrplan mit der Bildentstehung verbunden: Lichtausbreitung, Lichtquellen, Lichtbündel und Lichtstrahl, Lichtstrahl als Modellvorstellung, Verlauf spezieller Strahlen oder Lichtbündel an der Sammellinse, Brechungsverhalten paralleler Strahlen, Vereinfachung der Brechung an der Linsenmitte, Konstruktion der Bilder; Übersicht über Art und Lage der Bilder (vgl. ebd., 190 f.).
5. Die Bildentstehung bei der Sammellinse umfasst unterschiedliche Repräsentationen auf verschiedenen Abstraktionsebenen.
6. Nicht zuletzt sind für den Bereich der Strahlenoptik eine Reihe von domänenspezifischen Schülervorstellungen dokumentiert, die sich auch gerade auf das repräsentationale Verständnis der Schüler auswirken.
7. Um die kognitiven Anforderungen der experimentellen Durchführung des physikalischen Experiments zur Bildentstehung am Hohlspielgel bzw. der Sammellinse sowie der Datenauswertung und Interpretation zu erfassen, wurde eine kognitive Aufgabenanalyse durchgeführt (vgl. Gagné, Briggs & Wager, 1988). Im Zuge der Aufgabenanalyse wurden auch die verschiedenen Repräsentationsformen identifiziert, die bei der Durchführung, Auswertung und Interpretation des Schülerexperiments zur Bildentstehung am Hohlspiegel bzw. der Sammellinse zum Tragen kommen:
   - Die Ebene der verbalen Beschreibung der Phänomene (deskriptive Repräsentation, die sich an der konkreten Beobachtung orientiert).
   - Die Ebene der Messwerte. Die Schüler können Gegenstandsweite, Bildweite Gegenstandsgröße und Bildgröße abmessen und in tabellarischer Form festhalten (deskriptive Repräsentation, die durch die räumliche Anordnung eine depiktionale Komponente enthält) (vgl. Abbildung 8).

- Die Ebene der Modellbildung: Ausgehend vom Modell des Lichts als Strahl, kann die Entstehung der Bilder in einer Strahlenkonstruktion (depiktional-schematische Repräsentation) dargestellt werden (vgl. Abbildung 9).
- Die Ebene der generalisierenden Beschreibung der Verhältnisse von Gegenstandsweite, Bildweite und Abbildungsmaßstab. Diese verbale abstrakte Beschreibung stellt eine verbale Repräsentation dar (vgl. Abbildung 10), die auf die mathematische Repräsentation der Abbildungsgleichung hinführt (vgl. Abbildung 11).
- Die Ebene der Mathematisierung: hier das Abbildungsgesetz bzw. die Linsengleichung. Da der Lehrplan einen „sparsamen" Umgang mit der Mathematik nahelegt (vgl. ebd., S. 175), beschränkte sich die hier thematisierten mathematischen Zusammenhänge auf die Abbildungsgleichung (vgl. Abbildung 11). Ein Vorteil der Abbildungsgleichung besteht darin, dass ihre Herleitung geometrisch gut veranschaulicht werden kann, was ebenfalls mit den Zielen des Lehrplans (vgl. ebd., S. 190) konform ist (vgl. Abbildung 12).

Auf Basis der Aufgabenanalyse wurden die folgenden zehn kognitiven Schritte identifiziert. Die vollständige Aufgabenanalyse befindet sich in Anhang A (auf der Produktseite zu diesem Buch unter www.springer.com).

1. Aktivierung von Vorwissen und Nachdenken anregen. Die Aktivierung von Vorwissen bezieht sich in der 8. Klasse auf das Suchen von Alltagsbeispielen und technischen Anwendungen, wie etwa bei Autoscheinwerfern oder beim Kosmetikspiegel. Auf der repräsentationalen Ebene kommen hier Fotografien oder realistische Zeichnungen von Gegenständen ins Spiel.
2. Erste Fragestellungen identifizieren und Erklärungsversuche starten. Die Schüler können hier an bekannte Phänomene der Optik erinnert werden, wie das Reflexionsgesetz (Einfallswinkel = Ausfallswinkel). Sie operieren also mit abstrakten verbalen und schematischen Repräsentationen wie der geometrischen Anwendung des Reflexionsgesetzes.
3. Versuchsplanung der Schüler: Hypothesen aufstellen, Zusammenhänge auf Basis des Vorwissens und der Beobachtung vermuten und zu testende Einflussgrößen identifizieren.
4. Versuchsplan erarbeiten und einen Versuchsaufbau (vgl. Abbildung 7) festlegen: dabei die technische Umsetzung bedenken, den Versuchsaufbau sprachlich beschreiben. Zudem müssen die Schüler erkennen, was variiert, was beobachtet und was gemessen wird, d.h. abhängige und unabhängige Variablen müssen identifiziert und unterschieden werden.

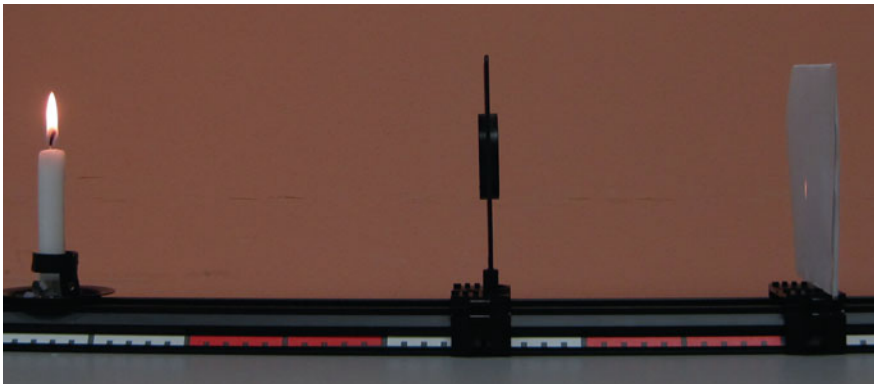


Abbildung 7: Experiment zur Bildentstehung bei der Sammellinse

5. Messwertetabelle planen (vgl. Abbildung 8): Festhalten, welche Größen gemessen werden soll und mit welche Ergebnissen zur Bestätigung oder dem Verwurf der Hypothesen in eine mathematisches Modell überführt werden kann.

Bildfall	Gegenstandsweite *g / cm*	Bildweite *b / cm*	Gegenstandsgröße *G / cm*	Bildgröße *B / cm*
gleich großes Bild	$g = 2f$	$b = 2f$	$G = B$	$B = G$
	10 cm	10 cm	4 cm	4 cm
verkleinertes Bild	$g > 2f$	$f < b < 2f$	$G > B$	$B < G$
	15 cm	7,5 cm	4 cm	2 cm
vergrößertes Bild	$f < g < 2f$	$b > 2f$	$G < B$	$B > G$
	8 cm	13,3 cm	4 cm	6,8 cm

Abbildung 8: Im Unterricht verwendete Messwertetabelle zur Bildentstehung bei der Sammellinse

6. Justierung des experimentellen Aufbaus und Messung der Werte unter Berücksichtigung des Versuchsplans. Hier wird erfordert, die konkrete Beobachtung zu quantifizieren und entsprechend in das Auswertungsschema (hier eine Tabelle, vgl. Abbildung 8) einzuordnen.
7. Auswertung und Abgleich mit Hypothese: Ggf. Fehlerquellen notieren und deren mögliche Ursachen erklären.
8. Modellbildung: Erstellung der Abbildungskonstruktion für die 3 Fälle. Vergrößertes Bild, verkleinertes Bild und gleichgroßes Bild. Die Modellbildung betrifft das Verstehen der schematischen Konstruktion des Strahlendiagramms (vgl. Abbildung 9).

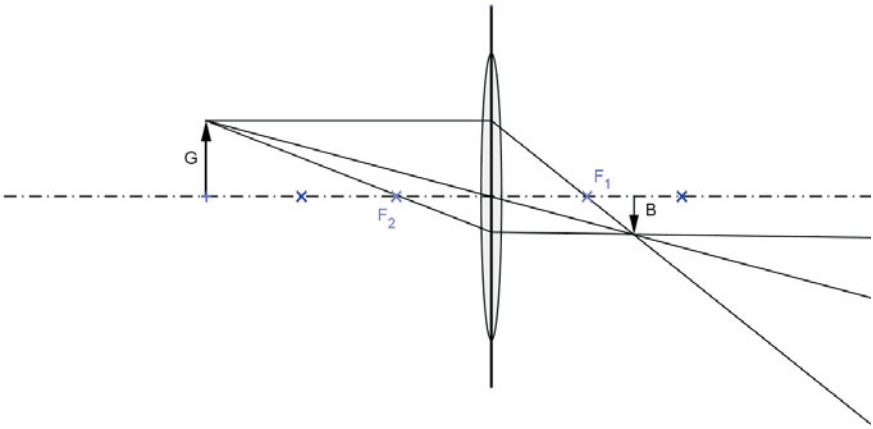


Abbildung 9: Strahlenkonstruktion zur Bildentstehung bei der Sammellinse

9. Bestätigung des Modells durch eine mathematische Erklärung: Dieser Schritt erfordert schließlich die schematische Darstellung des Strahlengangs (vermittelt über eine verbale Beschreibung) in eine mathematische Repräsentation zu übertragen, in diesem Fall in das Abbildungsgesetz.

Das Verhältnis von Bildgröße zu Gegenstandsgröße entspricht dem Verhältnis von Bildweite zu Gegenstandsweite.

Abbildung 10: Generalisierende Beschreibung der Verhältnisse von Gegenstandsweite, Bildweite und Abbildungsmaßstab

$$A = \frac{B}{G} = \frac{b}{g}$$

Abbildung 11: Abbildungsgleichung

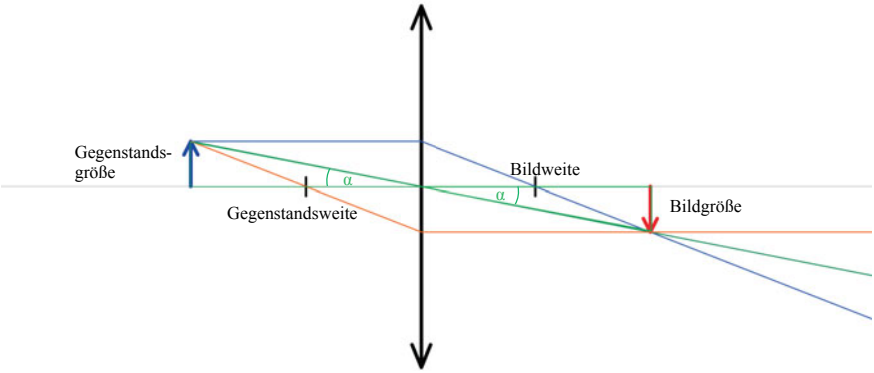


Abbildung 12: Geometrische Veranschaulichung der Abbildungsgleichung

10. Reflexion des Gelernten: Zusammenhänge der einzelnen Schritte reflektieren: Was war die Ausgangsfragestellung? In welche Teilziele konnte die Fragestellung zergliedert werden? Welche Operationen mussten angewandt werden, um diese Etappen zu erreichen, welche Repräsentationsformen wurden dabei genutzt? Wie gut wurden die Zwischenziele erreicht? In welcher Relation steht das Erreichte zur Ausgangsfragestellung? Konnte diese letztlich beantwortet werden?

2.1.2 Schülervorstellungen im Kontext der Bildentstehung bei der Sammellinse

Naive Vorstellungen im Bereich der Optik wurden insbesondere in der Fachdidaktik in den letzten Jahrzehnten genauer untersucht (vgl. Kärrqvist & Andersson, 1983; Guesne, 1985; Wiesner, 1986, 1992a, 1992b, 1994; Reiner et al., 2000). In der Regel wurden die Teilnehmer, Schüler oder Studenten, zu Alltagsphänomenen (z.B. „Was verstehst du unter Licht?“ zit. n. Guesne, 1985, S. 80) oder zu physikalischen Experimenten (vgl. Goldberg & McDermott, 1987, S. 109) befragt.

Generell ergaben die wissenschaftlichen Untersuchungen, dass die Teilnehmer die Konzepte nicht systematisch verwenden, dass viele Vorstellungen unter wissen-

schaftlichen Gesichtspunkten nicht akzeptabel waren und dass es schwierig ist, diese Vorstellungen durch übliche Instruktionen zu verändern. Solche naiven und wissenschaftlich unangemessenen Vorstellungen wurden dabei in den folgenden Bereichen gefunden: Entstehung von Licht und Schatten, Entstehung von Farben, physikalische Beschreibung des Sehvorgangs, Entstehung von Spiegelbildern und Entstehung reeller und virtueller Bilder bei der Sammellinse sowie ontologische Grundannahmen zum Thema Licht.

In dem folgenden Literaturrückblick werden naive Vorstellungen zu den Bereichen der Strahlenoptik vorgestellt, die in einem engeren inhaltlichen Zusammenhang des gewählten Lerninhalts der Bildentstehung bei der Sammellinse stehen. Diese betreffen alle genannten Bereiche außer der Entstehung von Licht und Schatten und der Entstehung von Farben.

Eine der ersten und maßgeblichen Studien zu naiven Vorstellungen über Licht von Jugendlichen hat Guesne (1985) durchgeführt. Sie interviewte 30 Jugendliche im Alter von 13 und 14 Jahren, die zuvor keinen Optikunterricht in der Schule erhalten hatten, in standardisierten Interviews.

Zu den wesentlichen Vorstellungen zählte, dass die Befragten Licht mit seiner Quelle, seinen Wirkungen oder einem Zustand gleichsetzten.

- Gleichsetzung mit der Quelle: Manche der interviewten Jugendliche lokalisierten das Licht ausschließlich in der Lichtquelle z.B. in der Glühbirne.
- Gleichsetzung mit den Wirkungen: Einige Kinder gaben an, Licht könne man nur an den Stellen sehen, auf denen das Licht auf der Wand helle Flecken erzeugt (z.B. durch Sonnenlicht oder durch die Reflexion eines Spiegels).
- Gleichsetzung mit einem Zustand: Jugendliche, die Licht mit einem Zustand gleichsetzen, bezeichnen Licht als Helligkeit, die sich z.B. mit dem Wetter ändere.

Als wichtigste Tatsache stellte sich hierbei heraus, dass kaum einer der befragten Jugendlichen annahm, gewöhnliche Gegenstände werfen Licht zurück. Guesne (1985) betont, dass diese Erkenntnis aus zwei Gründen von prinzipieller Bedeutung für den gesamten Bereich der Optik sei:

- Erstens sei es praktisch unmöglich, die Entstehung des Bildes von irgendeinem Gegenstand, der nicht selbst leuchtet z.B. in der Fotografie, zu verstehen, solange man diese Vorstellung nicht begriffen habe (vgl. Guesne, 1985, S. 85).
- Zweitens sei diese Annahme entscheidend dafür, den Sehvorgang zu begreifen. Da Kinder im Alltag Licht nur erkennen, wenn es einen deutlich wahrnehmbaren

> Effekt hervorbringe, glaubten die befragten Kinder nicht, dass auch bei nicht-selbst-leuchtenden Körpern Licht ins Auge gelange (vgl. Guesne, 1985, S. 91).

Diese Beobachtung und die zugehörige Deutung greift Wiesner (1992a, S. 16) nun wie folgt auf: Der Autor stellt ausgehend von den grundlegenden Lernschwierigkeiten einer unzureichenden physikalischen Sehvorstellung weitere zentrale Lernschwierigkeiten in der Strahlenoptik zusammen. Wiesner (1992a) bestätigte die Beobachtungen Guesnes (1985) und zeigte, dass Schüler auch nach dem Optikunterricht angaben, beleuchtete Gegenstände sehen zu können, ohne dass dazu Licht von dem wahrgenommenen Gegenstand ins Auge falle (vgl. ebd.).

Wiesner (1992a) sieht nun einen engen Zusammenhang zwischen einer fehlerhaften physikalischen Sehvorstellung und fehlerhaften Konzepten zur Streuung. Gerade das Konzept der Streuung betrifft den Kern einer solchen physikalischen Sehvorstellung, zu der der wichtige Aspekt zählt, dass beleuchtete Gegenstände selbst Licht abstrahlen (Wiesner, 1994, S. 7). So zeigte Wiesner (1986, S. 26), dass viele Schüler davon ausgehen, beleuchtete Gegenstände wie Tische, Bücher oder Bilder strahlten kein Licht ab. Auch hier fehlt nach Wiesner (1986) die Verbindung zwischen wahrgenommenen Gegenständen und den Augen des Betrachters. Daran schließt sich das Konzept an, das auftreffende Licht mache die Gegenstände hell, bleibe auf diesen liegen oder verschwinde allmählich (vgl. ebd.). Ebenfalls damit verbunden ist die Vorstellung, dass bei Lichtquellen mit geringer Intensität, wie Räucherstäbchen oder weit entfernte beleuchtete Fenster, kein Licht mehr ins Auge gelange (vgl. ebd.).

Guesne (1985) beschreibt im Zusammenhang mit der Vorstellung, beleuchtete Gegenstände strahlten kein Licht ab, eine Interpretation des Sehens, bei der dem Auge ein aktiver Part zugeschrieben wird, während dem Gegenstand nur eine passive Rolle zukommt. Dabei bleibe die Bewegung, die vom Auge zum Gegenstand hingehe, abstrakt. Das Subjekt werde dabei als Ursprung des Prozesses gesehen und nicht als Empfänger von Licht.

Die Schwierigkeiten mit der physikalischen Sehvorstellung schlagen sich nach Wiesner ebenfalls auf Lernschwierigkeiten mit dem Spiegelbild nieder. Auch hier erkennen viele Schüler nicht, dass das Licht aus der Richtung des Spiegels ins Auge fallen muss, damit das Spiegelbild wahrgenommen werden kann.

Nach Wiesner (1986, 1992a, 1992b) bereite Schülern aber nicht nur die Erklärung der Wahrnehmung, sondern auch die Lage des Spiegelbildes große Probleme: so werden die strahlengeometrische Konstruktion des Spiegelbildortes von den meisten Schülern als nicht überzeugend eingestuft. Die Schüler gingen oft davon aus, dass das Spiegelbild auf der Spiegeloberfläche liege (vgl. Wiesner,

1986, S. 26, 27, 1992a, 1992b, S 288). Der Spiegel werde auch oft als ein Gegenstand aufgefasst, der das Spiegelbild zum Betrachter zurückwerfe. Wiesner beschreibt hierzu folgende Situation (zit. n. Wiesner, 1992a):

> „Fragt man z.B. einen Schüler (oder auch einen Erwachsenen), was man tun kann, um in einem kleinen Taschenspiegel mehr vom eigenen Gesicht sehen zu können, hält die überwiegende Mehrzahl den Spiegel weiter weg vom Gesicht. […] Je weiter entfernt vom Gegenstand der Spiegel ist, desto kleiner erschienen nach dieser Meinung in dem Spiegel die Gegenstände, und dieses verkleinerte Bild wirft er zum Betrachter zurück" (S. 16).

Auch mit der Entstehung von reellen und virtuellen Bildern durch die Sammellinse ist eine Reihe von Fehlvorstellungen verbunden:

Schwierigkeiten bereiten Schülern insbesondere die physikalischen Vorstellungen des Abbildungsvorgangs. Viele Schüler nutzen zur Erklärung der Entstehung des reellen Bildes bei der Sammellinse nicht das Konzept einer Punkt-zu-Punkt-Abbildung.

Zu den gängigen Vorstellungen zählt: Das Bild ginge als Ganzes durch die Linse zum Schirm und werde dabei in der Linse umgedreht (Wiesner, 1994, S. 8). Der Autor bezeichnet dieses weitverbreitete Konzept als holistische Erklärung des Abbildungsvorgangs. Dass Lernende auf eine solche holistische Erklärung des Abbildungsvorgangs zurückgreifen, wird insbesondere bei Abdeckaufgaben deutlich. Unter der Annahme, das Bild werde als Ganzes vom Gegenstand aus durch die Linse auf den Schirm transportiert, ist es nur konsequent anzunehmen, dass ein Teil des Bildes abgeschnitten werde, wenn man eine Blende vor die Linse hält (vgl. Wiesner, 1992b, S. 288). Hält man eine ringförmige Blende vor die Linse, glauben viele Lernende entsprechend, das Bild werde ringförmig am äußeren Rand abgeschnitten. Wird die Linse zur Hälfte abgedeckt, gehen viele Schüler und auch Studenten davon aus, dass auch das reelle Bild zur Hälfte abgeschnitten werde, einige überlegen sich sogar, welche Hälfte des Bildes (obere versus untere Hälfte) betroffen sei (vgl. Goldberg & McDermott, 1987, S. 112; Wiesner, 1994, S. 8).

Goldberg und McDermott (1987) berichten in einer Studie über das Verständnis der Entstehung reeller Bilder durch die Sammellinse und den Hohlspiegel von der folgenden weiteren Verständnisschwierigkeit, die sie bei jungen Erwachsenen beobachten konnten. Die Autoren interviewten 80 College Studenten, die einen Einführungskurs in Physik besuchten, ca. die Hälfte aller Studenten hatten kein Vorwissen in Strahlenoptik durch Schule oder Universität erworben, die andere Hälfte hatte bereits ein experimentelles Praktikum in Optik absolviert. Den

Studenten wurde ein Versuchsaufbau gezeigt, bei dem eine Glühbirne, eine Linse und ein Schirm hintereinander auf einer optischen Bank montiert sind. Im Verlauf des Interviews wurden die Studenten gefragt, wo das Bild wäre, wenn man den Schirm entfernt und sie frei um den Versuchsaufbau im Raum herumgehen können. Nur wenige Studenten waren in der Lage zu erkennen, dass sich das Bild an der gleichen Position befindet wie der Schirm. Die übrigen Studenten gaben eine Erklärung ab wie etwa, das Bild sei auf oder in der Linse. Insbesondere war die Vorstellung verbreitet, dass ein Bild nur mit Hilfe eines Schirms gesehen werden kann und dass die Linse das Bild quasi einrahme (vgl. Goldberg & McDermott, 1987, S. 114).

Das virtuelle Bild bei der Sammellinse bereitet Schülern häufig ähnliche Schwierigkeiten wie die Entstehung des Spiegelbildes. Wiesner (1986) berichtet diesbezüglich zwei wesentliche Ansichten „(a) man schaut durch die Linse (quasi wie durch eine Gardine) hindurch auf eine Gegenstand und (b) das Bild liegt wie ein Spiegelbild auf der Linsenoberfläche“ (zit. n. Wiesner, 1986, S 28).

Weitere Erklärungen, die ebenfalls das Konzept der Punkt-zu-Punkt-Abbildung außer Acht ließen, bestehen in der Idee, die Linse konzentriere das Licht oder hinter der Linse sei mehr Licht bzw. seien mehr Strahlen vorhanden als vor der Linse (vgl. ebd., S. 158). Häufig werde die Entstehung reeller Bilder durch Spiegelung und Reflexion erklärt, dabei werde einem Gegenstandspunkt in der Regel nur ein Strahl zugeordnet und nicht ein divergierendes Strahlenbündel (vgl. ebd., S. 16).

Guesne (1985) befragte in ihrer Studie zu „Vorstellungen von Kindern über Licht“ auch zur Rolle der Sammellinse in der Funktion als Lupe und Brennglas. Dabei teilten sich die Kinder in zwei Antworttypen auf: Die erste Gruppe war der Ansicht, das Vergrößerungsglas mache das Licht größer, während die andere Hälfte davon überzeugt war, die Sammellinse konzentriere das Licht.

Kinder, welche das Konzept der Lichtkonzentration vertraten, waren der Ansicht, die gesamte Lichtmenge, die durch das Vergrößerungsglas hindurchgehe, bleibe hinter der Linse erhalten, was wissenschaftlich korrekt ist. Dass auch diese Gruppe von Kindern nicht notwendigerweise eine physikalisch angemessene Vorstellung von der Funktionsweise haben, zeigte sich in den angefertigten Zeichnungen der Kinder, mit denen die Kinder verdeutlichen sollten, wie die Linse das Licht konzentriere (vgl. Guesne, 1985, S. 87). So gab einer der befragten Jugendlichen an, ein einziger Strahl verlasse nach der Bündelung die Linse.

Kinder, die der Ansicht waren, das Vergrößerungsglas mache das Licht größer, gaben entweder an, hinter der Lupe sei mehr Licht als vor der Lupe oder das Licht werde hinter der Lupe verstärkt bzw. vermehrt. So stellte sich eines der befragten

Kinder vor, Licht könne nicht nur verstärkt werden, sondern auch verloren gehen oder verschwinden.

Gerade letztere Vorstellung zeigt, dass Kinder und auch Erwachsene mit wenig physikalischer Vorbildung wissenschaftstheoretisch grundlegend verschiedene Vorstellungen von Licht besitzen. Dies zeigt sich insbesondere auch in naiven Vorstellungen, die eine materialistische Konzeption von Licht annehmen. Reiner et al. (2000) beschreiben eine solche Konzeption unter Berufung auf Forschungsergebnisse verschiedener Studien (vgl. Andersson & Kärrqvist, 1983; Guesne, Séré & Tieberghien, 1983; Reiner, 1987; Smith, 1987; zit. n. Reiner et al., 2000, S. 14, 15) wie folgt: Wenn Lernende gefragt werden, wie das Sehen funktioniere, gaben sie an, dass Moleküle zwischen dem gesehenen Gegenstand z.B. einem „Buch“ und dem „Auge“ vorhanden seien. Der Sehvorgang wird demgemäß als das Ergebnis sich bewegender Lichtpartikel interpretiert. Hierbei wird Licht oft als Flüssigkeitsstrom beschrieben, der in Bewegung ist, sich aber auch in Ruhelage befinden kann.

Der eben beschriebene Literaturrückblick zeigt, dass naive und wissenschaftlich unangemessene Konzepte in allen Inhaltsbereichen, zu finden sind, welche für das Verständnis der Bildentstehung durch die Sammellinse relevant sind. Der nächste Abschnitt zielt darauf, diese Vorstellungen erstens auch vor dem Hintergrund der in Kapitel 1.3.2 beschriebenen Konzeptwechselansätzen einzuordnen und zweitens, mögliche Lernschwierigkeiten, die aus diesen Schülervorstellungen resultieren, beim Umgang mit (multiplen) Repräsentationen aufzuzeigen.

Zusammenfassend lassen sich die Schülervorstellungen diesen drei übergeordneten wissenschaftlich inadäquaten Konzepten zuordnen:

1. *Licht wird als eine Art Gegenstand oder Zustand aufgefasst*. Unter dieses Konzept lassen sich folgende Schülervorstellungen subsumieren:
   - Licht wird mit seiner Quelle gleichgesetzt.
   - Licht wird als Zustand („Helligkeit“) verstanden.
   - Licht wird als Substanz bzw. Materie (z.B. als eine Art Flüssigkeit) betrachtet und
   - Auftreffendes Licht macht die Gegenstände hell und bleibt auf den Gegenständen liegen.
2. *Reelles und virtuelles Bild werden als Gegenstände verstanden*. Diesem Konzept entsprechen die Vorstellungen:
   - Das Spiegelbild liegt auf der Spiegeloberfläche bzw. auf die Linsenoberfläche.
   - Das Bild geht als Ganzes durch die Sammellinse.

- Bei einer Lupe schaut man durch die Linse hindurch auf das Bild.
- Wird ein Teil der Linse abgedeckt, so wird auch ein Teil des Bildes abgeschnitten.

3. *Optische Geräte und auch das menschlichen Auge spielen eine aktive Rolle beim Sehvorgang oder bei der Bildentstehung.* Diesem Konzept lassen sich folgende Vorstellungen zuordnen:
   - Die Interpretation des Sehens, bei der dem Auge ein aktiver Part zugeschrieben wird, während dem Gegenstand nur eine passive Rolle zukommt.
   - Je weiter entfernt vom Gegenstand der Spiegel ist, desto kleiner erscheinen in dem Spiegel die Gegenstände und dieses verkleinerte Bild wirft der Spiegel zum Betrachter zurück.
   - Die Linse empfängt das Bild als Ganzes und dreht es aktiv um.
   - Die Linse konzentriert oder vergrößert das Licht.

Im folgenden Abschnitt wird der Versuch gestartet, die aufgeführten Schülervorstellungen vor dem Hintergrund des Rahmentheorieansatzes von Vosniadou (1992, 1994) und der Perspektive des fragmentarischen Wissens nach diSessa (1983, 1988, 1993) einzuordnen. An dieser Stelle soll ausdrücklich betont werden, dass die hier vorgenommene Einordnung ein hypothetisches Gedankenspiel darstellt. Ziel ist es, die Vorstellungen zu strukturieren und mögliche Erklärungen für das Zustandekommen der Vorstellungen zu finden. Zudem sollen Konsequenzen für das Lernen mit multiplen Repräsentationen aufgezeigt werden. Der Autorin liegen keine empirischen Daten vor, ob die hier skizzierten „möglichen" Rahmentheorien oder p-Prims in der formulierten Weise bei Lernenden beobachtet werden können. Die Frage, welcher Ansatz sich besser eignet, Schülervorstellungen in der Strahlenoptik zu erklären und ihre Überwindung zu fördern, ist ebenfalls kein Gegenstand dieser wissenschaftlichen Arbeit.

Fasst man bspw. das erste Konzept in Anlehnung an Vosniadou und Brewer (1992, 1994) bzw. Vosniadou (1992) als implizite epistemologische Rahmentheorien auf, könnten Probleme mit dem Konzept der Streuung und dem physikalischen Sehvorstellung wie folgt erklärt werden:

1. Ist Licht eine Art Substanz, dann strahlen beleuchtete Gegenstände kein Licht ab, weil das Licht auf diesen Gegenständen liegen bleibt und diese erhellt. Das Auge als aktiver Part würde diese Helligkeit erfassen.
2. Eine andere epistemologische Rahmentheorie (möglicherweise ein etwas weiter fortgeschrittener Ansatz) könnte in der Interpretation des Sehvorgangs

als das Ergebnis sich bewegender Lichtpartikel bestehen, wobei Licht oft als Flüssigkeitsstrom beschrieben wird, der zum Auge gelangt (hier käme dem Auge kein aktiver Part zu). Konsistent wäre diese Theorie jedoch nur dann, wenn der Strom von beleuchteten Gegenständen reflektiert würde.

3. Auch das Konzept, in welchem die Sammellinse eine aktive Funktion übernimmt, indem sie das Licht zu einem Strahl konzentriert, könnte als eine epistemologische Rahmenannahme aufgefasst werden. In diesem Fall würde das Konzept der „Punkt-zu-Punkt-Abbildung" keinen Sinn ergeben und einem Gegenstandspunkt würde dann (hinter der Linse) nur ein Strahl zugeordnet und nicht ein divergierendes Strahlenbündel.
4. Eine (möglicherweise etwas fortgeschrittenere) epistemologische Rahmentheorie zum Abbildungsvorgang könnte in der von Wiesner (1994) angesprochenen holistischen Konzeption des Abbildungsvorgangs bestehen, nach der das Bild als Ganzes durch die Linse zum Schirm geht und dabei in der Linse umgedreht wird (vgl. ebd.). Unter dieser Annahme ergibt sich bei einer partiellen Abdeckung, dass auch ein Teil des Bildes abgeschnitten werde. Ist die Blende ringförmig, wird das Bild entsprechend ringförmig am äußeren Rand abgeschnitten. Wird die Linse zur Hälfte abgedeckt, wird die obere bzw. die Hälfte des Bildes abgeschnitten.

Die Vermittlung wissenschaftlich adäquater Vorstellungen sollte darauf zielen, die Lernenden darin zu unterstützen ein internes mentales Modell der Lichtausbreitung und des Abbildungsvorgangs zu generieren. Auf diese Weise können die Lernenden die Zusammenhänge in kohärenter Weise intern dynamisch simulieren, z.B. indem sie sich vorstellen, was passiert, wenn man den Gegenstand von der Linse weg oder auf die Linse zuschiebt. Ist den Lernenden bewusst, dass von jedem Gegenstandspunkt eines leuchtenden Gegenstands Licht ausgeht, sollten sie erkennen können, dass bei einer Abdeckung der Linse das Bild lediglich schwächer bzw. blasser, aber dennoch vollständig abgebildet wird. Das mentale Modell der Lernenden würde als ein wesentliches Element die geometrische Strahlenkonstruktion beinhalten. Voraussetzung für ein wissenschaftlich angemessenes mentales Modell des Abbildungsvorgangs ist die physikalische Sehvorstellung. Die Schüler sollten entsprechend kognitiv dazu aktiviert werden, die Lichtstrahlausbreitung nachzuvollziehen und mögliche epistemologischen Rahmentheorien, wie das Konzept, Licht sei eine Art Substanz, die auf Gegenständen liege, zu revidieren. Auf Basis eines adäquaten mentalen Modells der Lichtausbreitung, sollten die Schüler in der Lage sein, die geometrischen Strahlenkonstruktionen in der Optik flexibel anzuwenden.

Fraglich ist jedoch, ob Lernende, die von Schülervorstellungen ausgehen, überhaupt über eine konsistente Erklärung des physikalischen Sehvorgangs oder der Bildentstehung verfügen. In Anlehnung an diSessa (1983, 1988, 1993) könnten die Lernschwierigkeiten auf bruchstückhafte intuitive Vorstellungen (sogenannte p-Prims) zurückgeführt werden. Zur Erinnerung, den Lernenden fehlt es gemäß dieser Theorie an den metakognitiven Fähigkeiten, die Inkonsistenz ihres Wissens zu erkennen.

Folgende mögliche p-Prims könnten auf Basis der erläuterten Schülervorstellungen formuliert werden.

1. Das Auge sieht, wenn es hell ist und kein Hindernis im Weg steht.
   - Dieses mögliche p-Prim beinhaltet die Vorstellung beleuchtete Gegenstände oder auch ein Spiegelbild könne man sehen, ohne dass dazu Licht ins Auge fallen müsse; bzw. beleuchtete Gegenstände strahlten kein Licht ab.
   - Auch die weitverbreitete Vorstellung, dass bei einer partiellen Abdeckung der Linse nur ein Teil des Bildes entsteht, lässt sich hier zuordnen. Da die Abdeckung ein „Hindernis“ darstellt, folgern einige Lernende, dass auch das Bild entsprechend abgeschnitten wird.
   - Auch die eher selten beschriebenen Vorstellungen einer aktiven Rolle des Auges könnten diesem möglichen p-Prim zugeordnet werde. Hierzu zählen die Vorstellungen von Sehstrahlen, die vom Auge ausgehend zum Gegenstand verlaufen oder die Vorstellung einiger Lernender, eine Katze könnte selbst bei völliger Dunkelheit noch Gegenstände sehen.
2. Licht erhellt Gegenstände.
   - Licht wird in folgender Schülervorstellung mit Helligkeit gleichgesetzt: Auftreffendes Licht mache die Gegenstände hell und bleibe auf ihnen liegen (wie eine Substanz). Von dieser Vorstellung gibt es zwei Varianten, die sich widersprechen: die Gleichsetzung von Licht mit seinen Wirkungen oder mit seiner Quelle.
3. Optische Geräte wie Spiegel oder Linsen erzeugen Bilder (als Ganzes), die den Gegenständen inhärent sind (z.B. auf der Linsen- oder Spiegeloberfläche liegen).
   - Dieses mögliche p-Prim könnte auf Alltagsbeobachtungen von Spiegelungen basieren. Die meisten dieser Spiegelungen sind nicht perfekt (wie z.B. bei Spiegelungen durch getönte Fensterscheiben an Gebäuden oder Autos oder Spiegelungen an Wasseroberflächen). Der Beobachter sieht ein Spiegelbild zusammen mit der sich spiegelnden Oberfläche und erhält den Eindruck, das Bild läge auf der Oberfläche. Nur bei sehr hochwertigen und großen

Spiegeln ist die Illusion eines Raumes hinter dem Spiegel glaubhaft. Liegt das Bild „auf dem Spiegel“, dann ist es auch nur folgerichtig, dass ein Befragter den Spiegel weiter weg halten will, wenn er oder sie einen größeren Bildausschnitt sehen möchte.

4. Je weiter entfernt ein Objekt ist, desto kleiner sieht man es (und einen desto größeren Ausschnitt kann man sehen).
   - Auch dieses mögliche p-Prim könnte direkt der Alltagserfahrung entspringen. So entfernt sich ein Fotograf, der eine Ganzkörper-Ansicht einer Person erstellen möchte, von der Person und sieht im Sucher seine Annahme bestätigt. Probleme ergeben sich nun, wenn diese Erfahrung auf Spiegelbilder übertragen wird. Dann liegt die Annahme nahe, den Spiegel weiter weg zu halten, um einen größeren Bildausschnitt sehen zu können.

Vorteile in der Einordnung der Schülervorstellungen in den Rahmentheorieansatz (vgl. Vosniadou & Brewer, 1992) liegen in der Erklärung der Fortschritte der Lernenden. Der Bezug der Vorstellungen zu impliziten ontologischen Hintergrundannahmen zeigt auf, dass die Lernschwierigkeiten der Schüler aus impliziten Annahmen resultieren, die Lernenden in der Regel nicht explizit zugänglich sind.

Mit diSessas Ansatz (1983, 1988, 1993) der Reorganisation von Wissensfragmenten lässt sich jedoch ebenfalls sehr gut erklären, dass die Schwierigkeiten der Lernenden aus Vorstellungen resultieren, die sie nicht explizit formulieren. Die Ursachen hierfür gründen sich nicht darauf, dass den Lernenden ihre epistemologischen Rahmentheorie nicht bewusst zugänglich sind (und sie diese daher auch nicht gezielt überprüfen können), sondern vielmehr darauf, dass die Vorstellungen fragmentarisch und inkonsistent sind und nicht explizit formuliert werden. Die Stärke dieses Ansatzes besteht darin, den diffusen Vorstellungen der Schüler keine vermeintlich höhere Konsistenz zu unterstellen, als sie tatsächlich aufweisen.

Geeignete Strategien für das Lernen mit multiplen Repräsentationen können darin bestehen, die Lernenden zu aktivieren, die eigene Vorstellungen, die extern repräsentiert wurden, mit Lösungen im Unterricht abzugleichen, wobei die Lernenden angehalten werden, die Lösungen nicht einfach zu übernehmen sondern auch kritisch zu prüfen. Indem die Lernenden eigene Repräsentationen, z.B. eine selbst erstellte Strahlenkonstruktion mit anderen Repräsentationen und den beobachteten Phänomenen abgleichen, können sie auf Inkonsistenzen aufmerksam gemacht werden, so z.B. bei der Darstellung der Entstehung von reellen Bildern bei Abdeckaufgaben. Auch diese Strategie zielt darauf, die Lernenden in der Konstruktion eines adäquaten mentalen Modells der Lichtausbreitung, des

Sehvorgangs und der Entstehung reeller Bilder zu unterstützen. Der Schwerpunkt unter der Perspektive des Lernens mit multiplen Repräsentationen bestünde darin, die unterschiedlichen Repräsentationen in konsistenter und kohärenter Weise zu verwenden. Ziel wäre es, ein Wissensnetz zu schaffen, in dem die unterschiedlichen Repräsentationen (wie die Messwertetabelle, die Strahlenkonstruktion und die Abbildungsgleichung) miteinander verknüpft werden.

2.2 Pilotstudie

2.2.1 Zielsetzung der Pilotstudie

Zahlreiche Forschungsergebnisse belegen, dass (naive) Schülervorstellungen sehr hartnäckig sind und sich gegenüber Veränderungsversuchen als widerständig erweisen (vgl. Goldberg & McDermott, 1987; Wiesner, 1992a; Vosniadou & Brewer, 1992; Tyson et al., 1997; Özdemir & Clark, 2007). Die naiven Vorstellungen der Lernenden beeinflussen das formale Erlernen wissenschaftlicher Konzepte und schlagen sich in externen Repräsentationen nieder (vgl. Cox, 1999), welche die Lernenden bei der Aufgabenbearbeitung erstellen. Der im ersten Teil der Arbeit dargestellte Forschungsstand zum Konzeptwechsel zeigt, dass die Veränderung des konzeptuellen Grundverständnisses der Schüler jeweils domänenspezifisch zu erreichen ist. Entsprechend gilt es, die Instruktionen auf die Schülervorstellungen zu beziehen, die in Relation zur Bildentstehung in der Strahlenoptik stehen. Die entwickelten Instruktionen zielen darauf ab, den Schülern zu zeigen, warum gegebenenfalls alternative oder qualitativ neue Repräsentationen zweckmäßiger zur Erklärung und Problemlösung sind (vgl. Tyson et al., 1997; Özdemir & Clark, 2007). Auf diese Weise sollten die Schülervorstellungen im Licht der neuen theoretischen Annahmen reinterpretiert werden. Die Pilotstudie verfolgte vorrangig zwei Hauptziele: erstens die Validierung von Erhebungsinstrumenten und zweitens die Validierung des Unterrichtsmaterials.

Die Treatmentbedingung beinhaltete Aufgabenstellungen, welche die Überwindung von gängigen Schülervorstellungen erfordern, die in einschlägiger Fachliteratur berichtet werden (Goldberg & McDermott, 1987; Wiesner, 1992; Reiner et al., 2000). In der Kontrollbedingung hingegen bearbeiteten die Schüler Aufgabenstellungen, die keine aktive Auseinandersetzung mit Schülervorstellungen erfordern. In beiden Bedingungen wurden die Schüler kognitiv aktiviert, sich mit verschiedenen Repräsentationsformen auseinander zu setzen.

2.2.2 Fragestellung und Hypothesen

Die folgenden Forschungsfragen wurden im Rahmen der Pilotstudie untersucht:

1. Führen die entwickelten kognitiv aktivierenden repräsentationalen Aufgaben zu einem Lernzuwachs in Bezug auf den Lerninhalt: die Bildentstehung in der Strahlenoptik?
2. Fördern kognitiv aktivierende Aufgaben, die darauf zielen, weitverbreitete Schülervorstellungen durch die Auseinandersetzung mit fachbezogenen Repräsentationen zu überwinden, Wissen, Problemlösen und den Umgang mit fachspezifischen Repräsentationen von Schülern in der Strahlenoptik?
3. Fördern diese Aufgaben das konzeptuelle Verständnis der Schüler?

Hypothese 1: Zur Beantwortung der ersten Forschungsfrage wurde analysiert, ob sowohl die Treatment- als auch die Kontrollgruppe Wissen und Problemlösen bei repräsentationsbezogenen Aufgaben und ihr konzeptuelles Verständnis in der Strahlenoptik nach der Intervention verbesserten.

Hypothesen 2 und 3: Zur Untersuchung der beiden Fragestellungen wurde geprüft, ob die Treatmentgruppe die Kontrollgruppe nach der Intervention in zwei Aspekten einen höheren Lernzuwachs erzielt.

- Erstens im Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen im Bereich der Strahlenoptik (erfasst in einem inhaltsspezifischen Leistungstest),
- Zweitens im konzeptuellen Verständnis im Bereich der Strahlenoptik (erfasst in einem Konzepttest zum Thema Strahlenoptik).

2.2.3 Stichprobe und Design

Die Stichprobe bestand aus 57 Schülern aus zwei Parallelklassen, welche die achte Klasse eines Gymnasiums besuchten. Eine Klasse wurde der Treatment- und eine Klasse der Kontrollbedingung zugeordnet. Alle Schüler waren im Alter von 13 und 14 Jahren. Der Stichprobe gehörten 24 Jungen und 33 Mädchen an. Die Teilnahme an der Studie war freiwillig und mit Schülern, Eltern und Lehrern im Vorhinein abgesprochen. Fünf Schüler fehlten entweder beim Prä- oder beim Posttest und wurden daher in der Analyse nicht berücksichtigt.

Bei der Pilotstudie handelte es sich um ein einfaktorielles quasi-experimentelles Prä-Posttest-Design. Die Treatmentbedingung beinhaltete Aufgabenstellungen, welche die Überwindung von gängigen Schülervorstellungen erfordern, die in einschlägiger Fachliteratur berichtet werden (Goldberg & McDermott, 1987; Wiesner, 1992a, 1992b; Reiner et al., 2000). In der Kontrollbedingung hingegen bearbeiteten die Schüler Aufgabenstellungen, die keine aktive Auseinandersetzung mit Schülervorstellungen erfordern. In beiden Bedingungen wurden die Schüler kognitiv aktiviert, sich mit verschiedenen Repräsentationsformen auseinander zu setzen. Es handelt sich um eine quasi-experimentelle Studie, da die Schüler im Klassenverbund als Gruppe den Bedingungen zugeteilt wurden. Grund für die Wahl eines quasi-experimentellen Designs bestand in der Einbindung in den regulären Schulalltag. Um den Stundenplan einzuhalten, konnten die Schüler nicht individuell den Bedingungen zugeordnet werden. In der Folge müssen Leistungsunterschiede, die von vornherein bestanden, in Kauf genommen werden. Verfahren der Parallelisierung auf Basis von Vorleistungen und / oder Intelligenz waren somit ausgeschlossen. Beide Klassen, Treatment- und Kontrollgruppe, wurden vom gleichen Lehrer im gleichen Zeitumfang von 135 min, also drei Schulstunden, unterrichtet. Die Prä- und Posttests fanden jeweils in der Physikstunde vor und nach der Unterrichtseinheit statt und erforderten jeweils insgesamt 45 min. (25 min. + 15 min.).

2.2.4 Durchführung und Unterrichtsmaterial

Um die Hauptstudie im Schuljahr 2010/2011 durchführen zu können, wurde die Pilotstudie Ende April und Anfang Mai 2010 umgesetzt. Zu diesem Zeitpunkt hatten die Schüler seit August 2009 Unterricht im Fach Physik. Das Thema Strahlenoptik war zwei Monate zuvor abgeschlossen worden, so dass die Schüler bereits Vorkenntnisse zum Thema Bildentstehung mit der Sammellinse besaßen. Aus diesem Grund fiel die Wahl des Lerngegenstands auf das Thema der Bildentstehung beim Hohlspiegel: Mit der Bildentstehung am Hohlspiegel wurde ein Lerninhalt gewählt, der in den repräsentationalen Anforderungen äquivalent zur Bildentstehung bei der Sammellinse ist und zugleich als Wiederauffrischung und Vertiefung früherer Lerninhalte in das Jahrescurriculum eingebunden werden konnte (Unterrichtsthema unmittelbar vor Beginn der Studie war Mechanik). Die Unterrichtsreihe umfasste drei Schulstunden zur Bildentstehung beim Hohlspiegel:

1. In der ersten Stunde erfolgte zunächst eine kurze Einführung mittels eines Demonstrationsexperimentes mit einer optischen Scheibe, bei der Brechungsvorgänge am Hohlspiegel demonstriert wurden. Nach der Einführung führten die Schüler ein Schülerexperiment zur Entstehung reeller Bilder am Hohlspiegel durch.
2. In der zweiten Stunde wurden die Strahlenkonstruktion und die Abbildungsgleichung behandelt.
3. In der dritten Stunde bearbeiteten die Schüler Übungsaufgaben zum Umgang mit verschiedenen Repräsentationsformen, wie der Strahlenkonstruktion und der Abbildungsgleichung, die es erforderten, diese Repräsentationsformen aufeinander und auf das Schülerexperiment zu beziehen

Die Treatmentgruppe bearbeitete kognitiv aktivierende Aufgaben, die verlangten, sich vertieft mit Repräsentationen auseinanderzusetzen und die in der Strahlenoptik bekannten weitverbreiteten Schülervorstellungen ansprechen. Im Gegensatz zur Treatmentgruppe bearbeitete die Kontrollgruppe ebenfalls kognitiv aktivierende Aufgaben, die es erforderten, sich vertieft mit Repräsentationen auseinanderzusetzen. Weit verbreitete Schülervorstellungen wurden jedoch nicht thematisiert. Ausführliche Informationen und weitere Beispiele zu den verwendeten Unterrichtsmaterialien finden sich in Hettmannsperger, Schnotz, Müller und Scheid (in Druck). Zur Verdeutlichung der Unterschiede zwischen Treatment- und Kontrollgruppe wird hier folgendes Aufgabenbeispiel genannt (vgl. Abbildung 13), indem die weit verbreitete Lernschwierigkeit thematisiert wird, dass sich die Lichtstrahlen, die ein reelles Bild formen, auch im freien Raum treffen können (vgl. Goldberg McDermott, 1987, S. 114).

Die Treatmentgruppe wurde gefragt, an welcher Stelle der Beobachter A, B, oder C ein scharfes Bild einer Kerze sehen kann, wenn man

a) einen intransparenten Schirm an der Position S aufstellt,
b) den intransparenten Schirm gegen einen transparenten Schirm austauscht oder
c) den Schirm entfernt.

Parallel zu dieser Aufgabe operierte die Kontrollgruppe mit derselben Repräsentation und wurde gefragt, an welcher Position ein Beobachter ein scharfes Bild der Kerze erkennen kann. Zudem wurde sie aufgefordert, den dargestellten Bildpunkt H der Kerzenflamme in der gegebenen Abbildung zu konstruieren. Die Kontrollgruppe wurde also kognitiv aktiviert, mit der gleichen Repräsentation zu operieren, die erwähnte Schülervorstellung wurde jedoch nicht thematisiert. Das

vollständige Unterrichtsmaterial und die zugehörige Unterrichtsplanung findet sich in Anhang B1 und B2 auf der Produktseite zu diesem Buch unter www.springer.com.

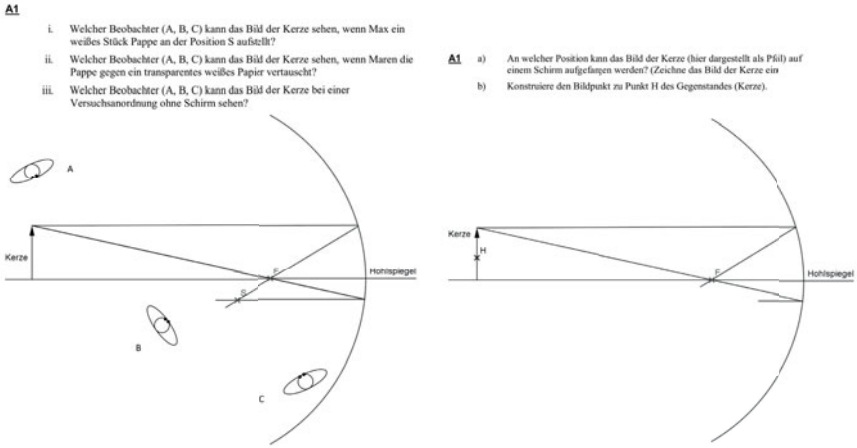


Abbildung 13: Aufgabenblatt 4: Übungen zur Strahlenkonstruktion und zur Abbildungsgleichung am Hohlspiegel, Aufgabe 1 [Material 7-TG] bzw. [Material 7-KG], links Aufgaben der Treatmentgruppe (TG), rechts Aufgabe der Kontrollgruppe (KG)

2.2.5 Variablen und Erhebungsinstrumente

2.2.5.1 Leistungstest

Die in einem Leistungstest erhobenen abhängigen Variablen Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen orientierten sich am Lehrplan für Gymnasien in Rheinland-Pfalz der Jahrgangsstufe 8. Der Test wurde von Jochen Scheid in Zusammenarbeit mit der Autorin eigens für die Studie entwickelt, da zum Zeitpunkt der Studie kein Instrument vorlag, welches das Themengebiet der Bildentstehung in Klassenstufe 8 hinsichtlich des Umgangs mit (multiplen) Repräsentationen erfasste. Der Test bestand aus neun Aufgaben und zielte darauf, Wissen und Problemlösen im Bereich der Strahlenoptik durch Aufgaben zu erfassen, die den Umgang mit fachspezifischen Repräsentationen erfordern. Die Testdauer betrug sowohl im Prä- als auch im Posttest 25 Minuten. Maximal konnten jeweils 17 Punkte je Test erreicht werden. Prä- und Posttest

enthielten die gleichen Aufgabenstellungen, wobei die Aufgaben des Prätests die Bildentstehung an der Sammellinse zum Thema hatten und die Aufgaben des Posttests die Bildentstehung am Hohlspiegel. Der vollständige Test findet sich in Anhang B3 auf der Produktseite zu diesem Buch unter www.springer.com.

Der Test umfasste folgende Aufgabentypen:

- Typ 1: drei Aufgaben, welche das Operieren mit Repräsentationen in unterschiedlichen Formaten erforderten (depiktionale und deskriptive Repräsentationen). Die repräsentationsbezogenen Aufgaben konnten nur korrekt gelöst werden, wenn das Denken der Schüler nicht durch spezifische Schülervorstellungen geleitet wurde (Aufgaben 1, 2a und 2b[12]). Aufgaben dieses Typs hatten zum Ziel, Wissen und Problemlösen beim Umgang mit Repräsentationen in unterschiedlichen Formaten im Hinblick auf die Überwindung spezifischer Schülervorstellungen zu erfassen.
- Typ 2: zwei Aufgaben, in denen die Schüler Repräsentationen in depiktional schematischer Form in eine deskriptiv verbale Repräsentationsform übersetzen sollten (Aufgaben 3 und 4) und Fehler in gegebenen depiktionalen Repräsentationen zu erkennen. Dieser Aufgabentyp zielte darauf, das fachliche Wissen im Umgang mit Repräsentationen und die repräsentationale Kohärenz zu erfassen.
- Typ 3: zwei Leistungstestaufgaben zur Abbildungsgleichung (Aufgaben 5a und 5b). Aufgaben des dritten Typs erfassten die Lernleistung der Schüler und prüften, ob die Schüler mit der mathematischen Repräsentation des Abbildungsvorgangs umgehen konnten. Mit Aufgabenstellungen, die prüften, ob die Schüler die Zweckmäßigkeit der mathematischen Repräsentation der Abbildungsgleichung erkennen, sollte auch eine metakognitiven Komponente von Repräsentationskompetenz erhoben werden.

Aufgaben des ersten Typs werden exemplarisch anhand der Aufgaben 2a und 2b erläutert (vgl. Abbildung 14 und Abbildung 15). Aufgabe 2 gliedert sich in zwei Teilaufgaben, die aufeinander aufbauen. Das Anwenden der Strahlenkonstruktion (Operieren mit der gegebenen Repräsentation) in Aufgabe 2a unterscheidet sich von einer Standardaufgabenstellung dadurch, dass der Schirm nicht an die richtige Stelle gerückt ist.

12 Die Aufgaben 2a und 2b wurden nur von der Autorin eingesetzt, da sie in spezifischer Weise darauf zielen, den Umgang mit Repräsentationen im Hinblick auf die Überwindung weitverbreiter Schülervorstellungen zu erfassen.

2. Aufgabe (4P)

Die Skizze zeigt den Gegenstand G und sein Bild B, das durch den Hohlspiegel entsteht. Der Bildpunkt Q zum Gegenstandspunkt P wurde richtig konstruiert. Die Strahlenkonstruktion ist nicht dargestellt. Der Schirm ist jedoch nicht an der richtigen Stelle aufgestellt, sondern ein Stück zum Spiegel hingerückt.

a) Zeige durch eine Zeichnung, dass auf dem Schirm anstatt des Bildpunktes Q ein unscharfer Bildfleck entsteht. Verwende hierfür die gegebene Abbildung (unten).

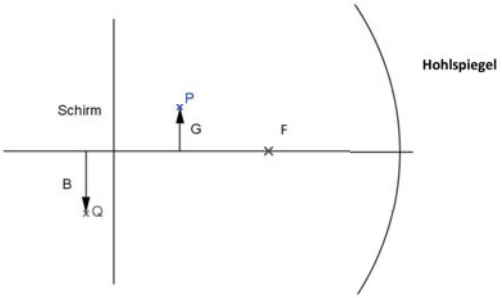


Abbildung 14: Aufgabenbeispiel aus dem Leistungsposttest: Anwenden der Strahlenkonstruktion

Diese Aufgabe prüft, ob die Schüler ein adäquates mentales Modell der geradlinigen Lichtausbreitung besitzen. Um die korrekte Größe des Lichtflecks zu bestimmen, müssen die Schüler neben den ausgezeichneten Strahlen auch den oberen Randstrahl einzeichnen, der gerade noch auf den Hohlspiegel trifft. Voraussetzung hierfür ist das Verständnis, das sich Licht in alle Richtungen ausbreitet. Dies berührt Schülervorstellungen zu den Konzepten der Lichtausbreitung und der Punkt-zu-Punkt-Abbildung. Aufbauend auf die Situation in Aufgabe 2a wurde im zweiten Aufgabenteil eine Blende vor den Hohlspiegel gesetzt.

Auch Aufgabe 2b zielt darauf, das Konzept-der Punkt-zu-Punkt-Abbildung zu prüfen und spricht die weitverbreitete Schülervorstellung an, eine Lochblende schneide das Bild ringförmig ab. Die Lösung bestand darin, auch hier diejenigen Strahlen einzuzeichnen, welche am Rand der Lochblende gerade noch auf den Hohlspiegel treffen. Über eine Transferleistung beim Operieren mit der bildlich-schematischen Repräsentation der Strahlenkonstruktion hinaus verlangt diese Aufgabe zudem, die bildliche Repräsentation in eine verbale Repräsentation zu übersetzen. Zudem prüft die Aufgabe die Kohärenz der verbalen und bildlichen Repräsentation. Eine Übersicht über die repräsentationalen Anforderungen der Aufgaben des Typs 2 und 3 finden sich in der Dissertationsschrift von Scheid (2013).

b) Wie ändert sich das Bild, wenn man jetzt noch eine Lochblende vor den Hohlspiegel stellt?

Begründe Deine Antwort durch eine Zeichnung. Verwende hierfür die gegebene Abbildung (unten).

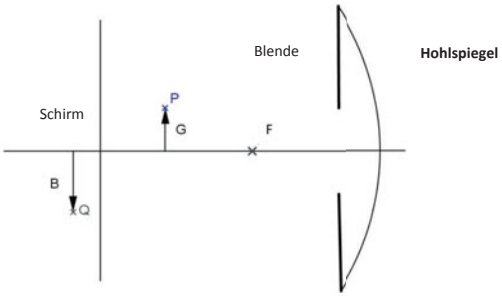


Abbildung 15: Aufgabenbeispiel aus dem Leistungsposttest: Lochblende vor Hohlspiegel

2.2.5.2 Konzepttest

Das konzeptuelle Verständnis der Schüler wurde in einem Konzepttest erfasst. Der Test wurde von der Autorin in Zusammenarbeit mit Jochen Scheid eigens für die Studie entwickelt, da zum Zeitpunkt der Studie kein Konzepttest zur Strahlenoptik vorhanden war. Ziel des Tests war die Erfassung physikalischer Grundkonzepte, welche allesamt für das Verständnis der Bildentstehung an der Sammellinse und am Hohlspiegel relevant sind.

Der Test beinhaltete Fragen zu folgenden Konzepten: Lichtausbreitung, Streuung, physikalische Sehvorstellung, zu ontologischen Konzepten, was unter „Licht“ verstanden werden kann, zur gerichtete Reflexion (Planspiegel) sowie zur Entstehung reeller Bilder bei der Sammellinse und am Hohlspiegel. Der als Multiple-Choice-Test konzipierte Konzepttest bestand aus 30 Items. Pro Aufgabe konnten maximal 2 Punkte erzielt werden, die maximal erreichbare Gesamtpunktzahl betrug somit 60. Die Fragen sollten erfassen, wie die Schüler mit bekannten Lernschwierigkeiten umgehen. Die Distraktoren bezogen sich hierbei auf bekannte Schülervorstellungen. Die Anzahl an Distraktoren orientierte sich je Aufgabe an der Anzahl alternativer Antwortmöglichkeiten, die auf Basis bekannter Schülervorstellungen gefunden werden konnten. Sie variiert zwischen 3 bis 6 Antwortmöglichkeiten.

Das folgende Aufgabenbeispiel (vgl. Abbildung 16) erfasste Konzepte zur Bildentstehung bei der Sammellinse. Korrekt war in diesem Fall die dritte Antwortmöglichkeit: „Lichtstrahlen eines Gegenstandpunktes werden durch die Sammellinse abgelenkt und treffen sich im Bildpunkt“, die Distraktoren bezogen sich auf die Schülervorstellungen:

- Das reelle Bild bei der Sammellinse entstehe durch Spiegelung oder Reflexion.
- Der holistischen Konzeption des Abbildungsvorgangs: Das Bild gehe als Ganzes durch die Linse zum Schirm und werde dabei in der Linse umgedreht.

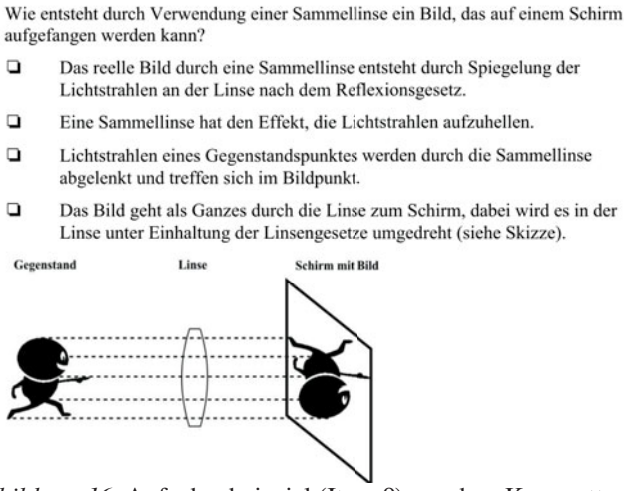


Abbildung 16: Aufgabenbeispiel (Item 9) aus dem Konzepttest

Pro Aufgabe war in der Regel eine Antwortmöglichkeit korrekt, mit Ausnahme der Aufgaben 4, 5, 7a, 8 und 14, in denen jeweils zwei Kreuze gesetzt werden mussten.

Die Antworten der Teilnehmer wurden mit zwei Punkten bewertet, wenn ausschließlich die korrekte Alternative gewählt wurde, ein Punkt wurde vergeben, wenn zusätzlich zu der richtigen Antwort auch ein Distraktor angekreuzt wurde (zwei Punkte für die Lösung weniger ein Punkt Abzug). In allen anderen Fällen wurde die Entscheidung mit null Punkten bewertet.

Im obigen Beispiel wurden also zwei Punkte vergeben, wenn die dritte Antwortmöglichkeit angekreuzt wurde. Die Schülerantwort wurde mit einem Punkt bewertet, wenn zusätzlich noch ein weiteres Kreuz gesetzt wurde. Eine Bewertung mit null Punkten erfolgte bei allen anderen Antwortverteilungen.

Bei Aufgaben, welche die Wahl von zwei Antwortalternativen erforderten, wurden zwei Punkte vergeben, wenn beide Lösungen gewählt wurden. Ein Punkt wurde

vergeben, wenn zusätzlich zu den beiden korrekten Antworten ein Distraktor für richtig befunden wurden (zwei Punkte für die Lösung weniger ein Punkt Abzug). Eine Bewertung mit null Punkten erfolgte bei allen anderen Antwortverteilungen. Folgende Grundkonzepte wurden in dem Test erfasst. Der vollständige Test befindet sich in Anhang B4 (auf der Produktseite zu diesem Buch unter www.springer.com).

2.2.6 Auswertung und Ergebnisse

2.2.6.1 Itemstatistiken zum Leistungstest

Um eine tragfähige Testfassung zu erstellen, wurde zunächst eine Itemanalyse des Leistungstests durchgeführt. Da sich die Werte von Treatment- und Kontrollgruppe weder im Prätest noch im Posttest wesentlich unterscheiden, werden an dieser Stelle die Werte für die Gesamtstichprobe (Treatment- und Kontrollgruppe) berichtet.

Die Analyse der Itemschwierigkeiten ergab, dass alle Items mit Ausnahme der Item 3 und 4 innerhalb des Toleranzbereichs von $0.20 \leq P_i \leq 0.80$ liegen (vgl. Anhang B5, Tabelle 1, auf der Produktseite zu diesem Buch unter www.springer.com). Die Aufgaben 3 und 4 wurden offenbar von vielen Schülern gut beantwortet, daher eignen sie sich offenbar wenig, die Leistung beim Umgang mit Repräsentationen differenziert zu messen. Item 3 und 4 bedürfen daher einer Überarbeitung.

Die Analyse der Trennschärfe zeigte, dass die korrigierten Trennschärfen der Items zwischen $.01 \leq r_{it} \leq .45$ für den Prätest und zwischen $-.11 \leq r_{it} \leq .35$ für den Posttest rangieren (vgl. Anhang B5, Tabelle 2, auf der Produktseite zu diesem Buch unter www.springer.com). Die Analyse belegt, dass Item 5a (Multiple Choice Item: Wozu kann man die Abbildungsgleichung gebrauchen?) durch eine negative Trennschärfe im Posttest auffiel. Dieses Item wird tendenziell sogar häufiger von Personen mit geringer Kompetenz in repräsentationsbezogenen Aufgaben korrekt beantwortet als von Personen mit hoher Kompetenz. Möglicherweise kommt dieses Ergebnis teilweise durch Raten zustande, da es sich um eine Multiple-Choice Frage handelt. Alternativ könnten auch die Distraktoren ungeeignet sein. Da dieses Item im Rahmen der Erfassung der Leistungen nicht als invertiert aufgefasst werden kann, wurde es von der Analyse ausgeschlossen.

Cronbachs Alpha erhöht sich unter Ausschluss des Items von $\alpha = .55$ auf $\alpha = .59$ im Posttest. Im Prätest sind die Werte nochmals niedriger ($\alpha = .49$ - Wert für alle Items, $\alpha = .50$ - unter Ausschluss von Item 5a). Für die Retest-Reliabili-

tät ergibt sich für die verbleibenden Items ein Wert von $r_{Test\text{-}Retest} = .50$. Ohne Berücksichtigung von Aufgabe 5a konnten nunmehr maximal 15.50 Punkte erzielt werden. Generell weisen die Items Trennschärfen unter $r_{it} = .40$ auf und auch die interne Konsistenz für den Posttest liegt mit knapp $\alpha = .59$ im unteren Bereich, was für eine notwendige und grundlegende Überarbeitung des Tests spricht.

Eine ausführlichere Diskussion der statistischen Kennwerte und der Struktur des Tests findet sich in der Dissertationsschrift von Scheid (2013)

2.2.6.2 Itemstatistiken zum Konzepttest

Um eine tragfähige Testfassung zu erstellen, wurde auch für den Konzepttest zunächst eine Itemanalyse durchgeführt.

Die Analyse der Itemschwierigkeit ergab, dass alle Items mit Ausnahme von Item 17a innerhalb des Toleranzbereichs von $0.20 \leq P_i \leq 0.80$ liegen (vgl. Anhang B5, Tabellen 3 und 4, auf der Produktseite zu diesem Buch unter www.springer.com).

Die Analyse der korrigierten Trennschärfen zeigt, dass die Mehrheit der Items in einem Trennschärfenbereich von $0.11 \leq r_{it} \leq 0.4$ im Posttest liegen. Bei den Items 3, 4, 5, 7a, 12a, 17a und 17b zeigen sich Werte nahe 0 oder sogar negative Werte (vgl. Anhang B5, Tabelle 5, auf der Produktseite zu diesem Buch unter www.springer.com).

Die Items 5, 7a und 17a und b, 12a wurden wegen der deutlich negativen Werte im Posttest ausgeschlossen. Eine Invertierung der Items lässt sich theoretisch nicht begründen: Schüler, welche die Bildentstehung bei der Sammellinse gut verstanden haben, sollten prinzipiell besser in der Lage sein diese Items zu beantworten als Schüler mit geringem Wissen.

Die Item 3 und 4, welche Trennschärfen nahe 0 aufweisen, wurden im Test belassen, müssen jedoch für die künftige Verwendung überarbeitet werden.

Für die korrekte Beantwortung der verbleibenden 26 Items konnten maximal 52 Punkte erzielt werden. Cronbachs Alpha erhöht sich unter Ausschluss der Items 5, 7b, 12a, und 17a von $\alpha = .72$ auf $\alpha = .78$ im Posttest. Im Prätest fällt Cronbachs Alpha entsprechend niedriger aus. Das geringere Wissen der Schüler im Vortest führt offenbar zur heterogeneren Beantwortung, was zu einem geringeren Wert der internen Konsistenz führt ($\alpha = .45$).[13] Für die Retest-Reliabilität ergibt sich für die verbleibenden 26 Items ein Wert von $r_{Test\text{-}Retest} = .39$.

13 Unter Ausschluss der Items 5, 7b, 12a und 17a ergibt sich der identische Wert von $\alpha = .45$.

2.2.6.3 Exkurs: Berechnung einer polychorischen Korrelationsmatrix für ordinalskalierte Daten

Da die Aufgaben des Konzepttests auf Itemebene mit den Abstufung 0, 1 oder 2 Punkte ordinalskaliert sind, wurde der Berechnung der Trennschärfen und Cronbachs Alpha in der Pilotstudie und in der später dargestellten Hauptstudie eine polychorische Korrelationsmatrix für ordinalskalierte Daten zugrunde gelegt.

Das Modell der polychorischen bzw. tetrachorischen Korrelation basiert auf der Annahme, dass eine latente kontinuierliche Variable durch Schwellenparameter in zwei bzw. drei oder mehr Kategorien eingeteilt wird. Je nachdem, wo der Wert einer Versuchsperson liegt, wählt die jeweilige Person eine der Kategorien. Die Schwellenparameter werden hierbei auf Basis der Häufigkeiten der Kategorien bestimmt. Unter der Annahme, dass nicht nur jede der latenten Variablen einer Normalverteilung folgt, sondern auch beide latente Variablen bivariat normalverteilt sind, lässt sich nun die Produkt-Moment-Korrelation der beiden latenten Variablen schätzen (vgl. Eid, Gollwitzer & Schmitt, 2011, S. 516). Die Berechnung basiert im Fall der tetrachorischen Korrelation, die ein Sonderfall der polychorischen Korrelation darstellt, darauf, dass sich der tetrachorische Koeffizient r_{tet} als Kosinus eines Winkels beschreiben lässt (vgl. ebd., S. 529):

$$r_{tet} = \cos\left(\frac{180\,^\circ}{1+\sqrt{\frac{n_{11}n_{22}}{n_{12}n_{21}}}}\right)$$

$n_{11}, n_{12}, n_{21}, n_{22}$ = Anzahlen in den jeweiligen Merkmalskombinationen, welche man durch eine Kreuztabelle darstellen kann

Der Berechnung liegt allgemein folgender Ansatz zugrunde: Gegeben seien zwei beobachtete ordinalskalierte Variablen *A* und *B* (mit in unserem Falle jeweils drei möglichen Werten). Es wird davon ausgegangen, dass A und B nach einem System von Grenzen x_1, x_2, y_1, y_2 wie folgt von unbeobachteten, metrischen Variablen X und Y abgeleitet werden können:

$$A = \begin{cases} 0 & falls\ X < x_1 \\ 1 & falls\ x_1 < X < x_2 \\ 2 & falls\ x_2 < X \end{cases}$$

$$B = \begin{cases} 0 & falls\ Y < y_1 \\ 1 & falls\ y_1 < Y < y_2 \\ 2 & falls\ y_2 < Y \end{cases}$$

Weiterhin wird angenommen, dass *X* und *Y* multivariat normalverteilt sind.

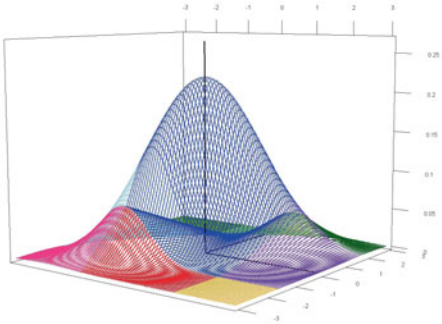
Ziel ist es, aus der Kontingenztafel von *A* und *B* die Korrelation der zugrundeliegenden Variablen X und Y zu schätzen. Hierzu wird die Maximum-Likelihood Methode angewendet. Freie Parameter sind die Korrelation (ϱ_{XY}) und die Grenzen (x_1, x_2, y_1, y_2). Mittelwerte und Standardabweichungen können (μ_X, σ_X) = (0,1) und (μ_Y, σ_Y) = (0,1) gewählt werden.

Die Berechnung der polychorischen Korrelationen erfolgte mit dem R-Befehl „polychor“ des Paketes „polycor“. Nach Olsson (1979) bieten sich folgende Methoden zur Lösung des Problems, welche beide im R-Befehl polychor aufrufbar sind:

1. Es wird eine vollständige Maximum-Likelihood Maximierung durchgeführt, bei welcher ϱ_{XY} und die Grenzen gleichzeitig geschätzt werden.
2. Es werden zuerst die Grenzen aus den marginalen relativen Häufigkeiten mittels der Inversen der Normalverteilung berechnet. Danach wird ϱ_{XY} mittels Maximum Likelihood Schätzung bestimmt. Diese „Two-Step-Method“ ist schneller und numerisch einfacher als die erste Methode, wenngleich erstere formal korrekter ist.

Im Folgenden ist dies beispielhaft dargestellt. Zu sehen ist die gemeinsame Dichte zweier multivariat normalverteilter Zufallsvariablen, welche jeweils eine Standardnormalverteilung als Randverteilung haben und mit $\varrho = 0.7$ korreliert sind. Die Farben repräsentieren die 3 x 3 Felder der Kontingenztafel der ordinalen Variablen. Das Integral über die einzelnen Farben entspricht der relativen Häufigkeit der ordinalen Kombination.

Da für die erste Methode Konvergenz nicht immer gegeben war, basieren die im Rahmen dieser Arbeit vorgestellten Ergebnisse (Maße der internen Konsistenz als Schätzer für die Realiblität, Korrelationsmatrizen) auf der zweiten Methode.



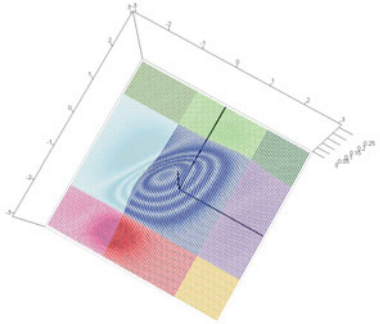


Abbildung 17 und *Abbildung 18*: Gemeinsame Dichte zweier multivariat normalverteilter Zufallsvariablen aus unterschiedlichen Perspektiven

2.2.6.4 Erste Hypothese: Veränderungen durch die Intervention

Zunächst wurde untersucht, ob die vorliegenden kognitiv aktivierenden repräsentationalen Aufgaben dazu geeignet sind, den Lerninhalt der Bildentstehung am Hohlspiegel inklusive der hiermit verbundenen physikalischen Grundkonzepte zu vermitteln.

Hierzu wurde analysiert, ob durch die eingesetzten Aufgaben überhaupt eine Änderung des Lernzuwachses prä – post erzielt werden kann. Wie Tabelle 1 zeigt, schneiden alle teilnehmenden Schüler sowohl im Leistungs- als auch im Konzeptposttest besser ab als in den jeweiligen Prätests.

Tabelle 1 Deskriptive Ergebnisse Mittelwerte und Standardabweichungen für Treatment- (TG) und Kontrollgruppe (KG)

AV[a]	Zeitpunkt	*TG (n = 27)*		*KG (n = 25)*		Alle *(N = 52)*	
		M	*SD*	*M*	*SD*	*M*	*SD*
LT[b]	prä	5.00	1.81	7.64	2.11	6.26	2.35
	post	9.50	3.63	11.40	1.72	10.41	3.01
KT[c]	prä	15.41	4.98	17.08	5.75	16.21	5.38
	post	27.74	10.34	22.44	5.98	25.19	8.86

[a]AV: abhängige Variable
[b]LT: Leistung bei repräsentationsbezogenen Aufgaben, maximal erreichbare Punkte: 15.50
[c]KT: Konzeptuelles Verständnis; maximal erreichbare Punkte: 52

Zur Überprüfung, ob diese Unterschiede auch signifikant sind, wurde ein t-Test für abhängige Stichproben durchgeführt, wobei jeweils die Ergebnisse des Prätests mit den Ergebnissen des Posttests verglichen wurden.

H_0 : $\mu_{d(post\text{-}präl)} \leq 0$
H_1 : $\mu_{d(post\text{-}präl)} > 0$

Voraussetzung hierfür ist die Normalverteilung für die jeweiligen Mittelwertdifferenzen (im Leistungstest und im Konzepttest). Diese Voraussetzung wurde per QQ-Plot und per Shapiro-Wilk-Test überprüft (vgl. Nachtigall & Wirtz, 2009, S. 171). Sie war für beide Tests, in der Treatmentgruppe, in der Kontrollgruppe wie auch in der Gesamtstichprobe gegeben.

Die Ergebnisse in Gemäß Cohen (1988, zitiert nach ebd.) handelt es sich in beiden Fällen von *d''* um einen großen Effekt.

Tabelle 2 bestätigen, dass der Mittelwertevergleich vor und nach der Unterrichtseinheit mit Ausnahme des Unterschieds im Konzepttest für die Kontrollgruppe auf dem $\alpha = 0.001$ Niveau signifikant sind und zwar sowohl was die Ergebnisse im Leistungstest $t(51) = 10.87$, $p < .001$, $d'' = 1.51$ als auch, was das Abschneiden im Konzepttest angeht, $t(51) = 6.21$, $p < .001$, $d'' = 1.08$. Die Ergebnisse unterstützen somit die erste Hypothese. Bei *d''* handelt es sich um die aus den Daten geschätzte Effektstärke für den Effekt δ'' auf Ebene der Population. Die Berechnung von δ'' errechnet sich nach Eid et al. (2011, S. 353) wie folgt:

$$\delta'' = \frac{\mu_D}{\sigma_D}$$

μ_D: Mittelwert der Differenz in der Population
σ_D: Standardabweichung der Differenz in der Population

Gemäß Cohen (1988, zitiert nach ebd.) handelt es sich in beiden Fällen von *d''* um einen großen Effekt.

Tabelle 2 Ergebnisse Vergleich des Mittelwertevergleichs für abhängige Stichproben prä und post

AV[c]	*TG[a] (n = 27)*			*KG[b] (n = 25)*			Alle *(N = 52)*		
	M (SD)	*t(26)*	*p*	*M (SD)*	*t(24)*	*p*	*M (SD)*	*t(51)*	*p*
LT[d]	4.50 (3.25)	7.19	<.001	3.76 (2.07)	9.07	<.001	4.14 (2.75)	10.87	<.001
KT[e]	12.33 (8.85)	7.24	<.001	5.36 (6.21)	4.32	<.001	8.98 (8.28)	6.21	<.001

[a]TG: Treatmentgruppe
[b]KG: Kontrollgruppe
[c]AV: abhängige Variable
[d]LT: Leistung bei repräsentationsbezogenen Aufgaben
[e]KT: Konzeptuelles Verständnis

2.2.6.5 Zweite Hypothese: Wirkung des Treatments auf die Physikleistung

Zur Überprüfung von Hypothese 2 wurde analysiert, ob die Treatmentgruppe die Kontrollgruppe hinsichtlich Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen (erfasst durch den Leistungstest) übertrifft.

Die Aufgaben des Präleistungstests unterscheiden sich von den Aufgaben des Postleistungstests in einem Punkt: zwar bearbeiteten die Schüler parallele Fragestellungen, doch im Prätest bezogen sich die Aufgaben mit den gleichen Formulierungen der Fragestellungen jedoch auf die Bildentstehung bei der Sammellinse, während im Posttest die Bildentstehung beim Hohlspiegel geprüft wurde. Diese Entscheidung begründete sich durch die Zielsetzung, das Vorwissen der Schüler möglichst genau einzuschätzen. Die Verwendung von Aufgaben zur Bildentstehung bei der Sammellinse bot sich an, weil die Schüler bereits Vorkenntnisse hierzu im Unterricht erworben hatten. Die Verwendung einer Varianzanalyse mit Messwiederholung war hiermit jedoch ausgeschlossen.

Zur Prüfung der zweiten Hypothese wurde eine ANCOVA als Analysemethode gewählt, da ein t-Test für unabhängige Stichproben im Leistungsvortest ergab, dass die Kontrollgruppe von einem höheren Niveau aus startete als die Treatmentgruppe $t(50) = 3.96, p < .001$. Formal kann das getestete Hypothesenpaar wie folgt dargestellt werden.

- H_0 : $\mu_{postTG} - \mu_{postKG} \leq 0$
- H_1 : $\mu_{postTG} - \mu_{postKG} > 0$

Die Kovarianzanalyse erfordert die Erfüllung folgender Voraussetzungen (Bortz, 2005, S. 369): erstens Intervallskaliertheit der Daten, zweitens Normalverteilung der Daten, drittens Varianzhomogenität und viertens die Korrelation der Kovariate mit der abhängigen Variablen.[14]

Zu 1): Auf Basis der Testkonstruktion wurde von der Intervallskaliertheit der Daten für die Bildung der Gesamtpunktzahl ausgegangen.

Zu 2): Zur Prüfung der Normalverteilungsannahme wurde jeweils für den Prä- und für den Posttest ein Shapiro-Wilk-Test durchgeführt und ein QQ-Plot erstellt (vgl. Nachtigall & Wirtz, 2004, S. 171).[15] Im Prätest zeigte sowohl der QQ-Plot als auch der Shapiro-Wilk-Test, dass die Daten hinreichend einer Normalverteilung ähneln. Dies gilt jedoch nicht für den Posttest: Hier handelt es sich um eine rechtssteile und linksschiefe Verteilung. Gemäß Bortz (2005b, S. 287) können Verletzungen der Normalverteilungsannahme bei schiefen Verteilungen und Stichproben $N > 10$ in Kauf genommen werden.

Zu 3): Der Levene-Test auf Gleichheit der Fehlervarianzen ergab ein signifikantes Ergebnis, so dass auch die Voraussetzung der Homogenität der Varianzen verletzt ist ($F(1, 50) = 16.63$, $p < .001$). Bei gleich großen Stichproben wird der F-Test durch heterogene Varianzen jedoch nur unerheblich beeinflusst (vgl. Bortz, 2005, S. 287). Bei der Interpretation der Ergebnisse wurden die entsprechenden Korrekturen, die SPSS ausgibt, berücksichtigt.

Zu 4): Abhängige Variable und Kovariate korrelieren signifikant miteinander ($r(50) = .50, p < .001$). Beim Vergleich der Ergebnisse von ANOVA und ANCOVA zeigt sich, dass die Berücksichtigung der Kovariate zugunsten der Treatmentbedingung ausfällt. Es ergibt sich ein positives Regressionsgewicht für die Kovariate, $b = 0.60$, $SD = 0.19$; $t = 3.10$, $p = .003$.

Insgesamt zeigte sich in der ANCOVA kein signifikanter Unterschied zwischen Treatment- und Kontrollgruppe im Posttest, $F(1, 52) = 1.36$, $MSE = 0.96$, $p = .714$ *n.s.*[16].

14 Kovariaten mit geringer Reliabilität reduzieren die Teststärke und können bei nicht randomisierten Untersuchungen (z.B. in quasi-experimentellen Studie wie dieser) zu Verzerrungen führen (vgl. ebd.).

15 Dieser Test bietet sich bei kleinen Stichproben an.

16 n.s. = nicht signifikant

2.2.6.6 Dritte Hypothese: Wirkung des Treatments auf das konzeptuelle Verständnis

Schließlich wurde geprüft, ob die Treatmentgruppe im Konzepttest einen höheren Lernzuwachs verzeichnen kann als die Kontrollgruppe. Da der Kenntnisstand von Teilnehmern unter verschiedenen Bedingungen wiederholt mit dem gleichen Testinstrument erhoben wurde, bot sich als Analysemethode eine Varianzanalyse mit Messwiederholung an.

Zur Analyse der Ausgangsbedingungen ergab ein t-Test für unabhängige Stichproben im Konzeptvortest keine signifikanten Mittelwertunterschiede zwischen den beiden Gruppen ($t(50) = 1.12$, $p = .267$, n.s.).

An die Durchführung einer Varianzanalyse mit Messwiederholung sind zunächst die gleichen Voraussetzungen (1 bis 3) wie an eine ANOVA gebunden. Eine weitere Voraussetzung für die Durchführung einer zweifaktoriellen Varianzanalyse mit Messwiederholung ist Sphärizität. Sphärizität ist dann gegeben, wenn die Varianz der Differenzen (Kovarianz) für jedes beliebige Paar von Bedingungen gleich ist, oder anders ausgedrückt, wenn keine beliebigen zwei Bedingungen mehr oder weniger abhängig voneinander sind als irgendwelche anderen beiden. Sphärizität wird also erst ab drei Vergleichen zu einem möglichen Problem.

Zu 1) Auf Basis der Testkonstruktion wurde auch hier von der Intervallskaliertheit der Daten für die Bildung der Gesamtpunktzahl ausgegangen.

Zu 2) Die Prüfung der Normalverteilungsannahme erfolgte jeweils für den Prä- und für den Posttest per Shapiro-Wilk-Test und per QQ-Plot (vgl. Nachtigall & Wirtz, 2004, S. 171).[17] Im Prä- und im Posttest zeigte sich sowohl im QQ-Plot wie auch im Shapiro-Wilk-Test, dass die Daten hinreichend einer Normalverteilung ähneln.

Zu 3) Der Levene-Test auf Gleichheit der Fehlervarianzen ergab kein signifikantes Ergebnis, so dass von der die Homogenität der Varianzen ausgegangen werden kann.

Zu 4) Die im Mauchly-Test auf Sphärizität angegebene Signifikanz von $p < .001$ ist kleiner als das Signifikanzniveau von $p < .05$, was darauf hinweist, dass für die Daten keine Sphärizität vorliegt. Da jedoch nur zwei Gruppen miteinander verglichen werden, sind die Ergebnisse unter der Annahme, Sphärizität sei gegeben, identisch mit den Ergebnissen welche auch unter Berücksichtigung der Korrekturen von Box (zit. n. Rasch, Friese & Naumann, 2010, S. 111), von Greenhouse & Geisser (1959) (zit. n. ebd.) sowie von Huynh & Feldt (1976) (zit.n.

17 Dieser Test bietet sich bei kleinen Stichproben an.

Field, 2009, S. 461; vgl. auch Rasch et al., 2006, S. 111) ausgegeben werden ($\varepsilon_{box=}$ 1; $\varepsilon_{Huynh\text{-}Feldt} = 1$; $\varepsilon_{Greenhouse\text{-}Geisser} = 1$).

Die Analyse wurde, wie folgt, umgesetzt: Die Quadratsummenanteile zwischen den Personen bezogen sich auf die Unterscheidung der Bedingung Treatment- versus Kontrollgruppe und die Quadratsummenanteile innerhalb der Personen auf den Faktor Zeit (Prä- versus Posttest). Für die Klärung der Fragestellung war insbesondere der Interaktionseffekt: Bedingung x Zeit relevant. Zur Analyse bot sich ein „einfacher" linearer Kontrast an, bei dem der Effekt jeder Faktorstufe mit dem Effekt einer Referenzfaktorstufe verglichen wird (vgl. Rudolf & Müller, 2004, S. 100). Für diese Analyse wurde der Faktor Zeit (Prämessung) mit dem Effekt der folgenden Faktorstufe (Postmessung) verglichen. Formal kann das getestete Hypothesenpaar wie folgt dargestellt werden.

$$H_0: \quad (\mu_{postTG} - \mu_{präTG}) - (\mu_{postKG} - \mu_{präKG}) \leq 0$$
$$H_1: \quad (\mu_{postTG} - \mu_{präTG}) - (\mu_{postKG} - \mu_{präKG}) > 0$$

Die ANOVA mit Messwiederholung belegt einen signifikanten Interaktionseffekt zwischen Zeit und Bedingung $F(1, 50) = 10.65$, $MSE = 315.60$, $p = .002$, $\eta_p^2 = .18$.

Abbildung 19 zeigt, dass die Treatmentgruppe einen signifikant höheren Lernzuwachs erreichte als die Kontrollgruppe. Die Effektstärke wurde mit dem Programm: OMEGA 3.1 (Fischer, 2001) berechnet. Auf Ebene der Population errechnet sich wie folgt (Rasch et al., 2010, S. 37):

$\sigma^2_{systematisch}$	systematische Varianz auf Ebene der Population
σ^2_{Gesamt}	Gesamtvarianz auf Ebene der Population

Der auf Basis der zugrundeliegenden Stichprobe *geschätzte* Effekt ω^2 für den Effekt in der Population betrug: $\omega^2 = .16$. Gemäß Cohen (1988) handelt es sich um einen starken Effekt: 16 % der erklärten Varianz lassen sich auf das Treatment zurückführen. Die dritte Hypothese konnte daher bekräftigt werden.

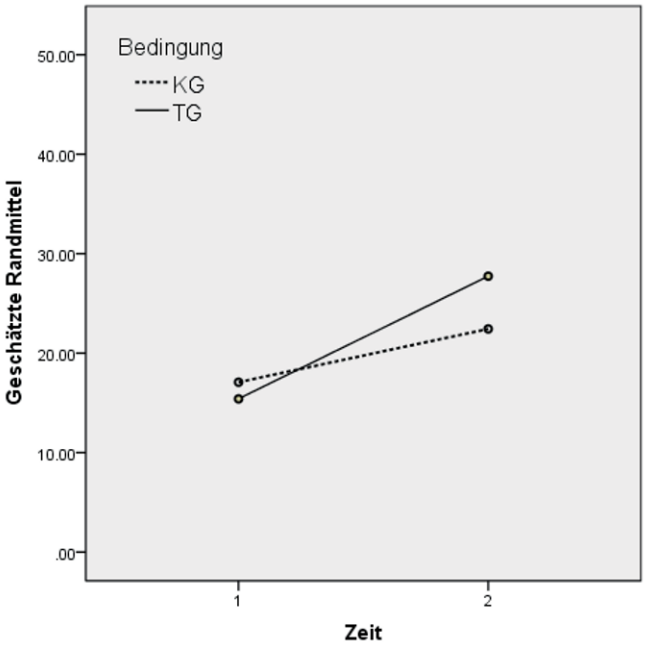


Abbildung 19: Interaktion zwischen Bedingung Treatment- (TG) versus Kontrollgruppe (KG) und Zeit zur Veranschaulichung des höheren Lernzuwachses der TG in Bezug auf das konzeptuelle Verständnis

2.2.7 Diskussion der Pilotstudie und Konsequenzen für die Hauptstudie

Das Hauptziel der Pilotstudie bestand darin, die entwickelten Lernmaterialien und Tests in der Praxis zu erproben. Die Ergebnisse weisen darauf hin, dass die entwickelten Materialien sowohl in der Treatment- als auch in der Kontrollgruppe lernwirksam sind.

Die Pilotstudie zeigt zudem, dass eine relativ kurze Intervention (135 min.), welche auf die Überwindung von weitverbreiteten Schülervorstellungen abzielt, zu einer signifikanten und praktisch bedeutsamen Verbesserung des konzeptuellen Verständnisses in der Strahlenoptik führt. Durch die Thematisierung von Schülervorstellungen durch Aufgaben, welche Schüler kognitiv aktivieren, sich mit multiplen Repräsentationen auseinanderzusetzen, kann offenbar das konzeptuelle Verständnis der Schüler gefördert werden. Der höhere Lernzuwachs in der Treatmentgruppe weist darauf hin, dass sich die Schüler signifikant häufiger

für korrekte Antworten und gegen Antwortalternativen entschieden haben, welche weitverbreitete Schülervorstellungen beinhalten. Angesichts der Hartnäckigkeit von Schülervorstellungen lässt sich dies durchaus als Erfolg werten.

Im Leistungstest bestehen Unterschiede im Vorwissen zwischen beiden Gruppen: So startete die Kontrollgruppe mit besseren Lernvoraussetzungen als die Treatmentgruppe. Unter Berücksichtigung der Gesamtpunktzahl im Leistungsvortest als Kovariate in der Analyse konnte kein signifikant höherer Lernzuwachs in der Treatment- im Vergleich zur Kontrollgruppe festgestellt werden.

Die Ergebnisse sind aus fünf Gründen jedoch mit Vorsicht zu interpretieren.

1. Für die Ergebnisse des Leistungstests besteht eine maßgebliche Einschränkung in der geringen Reliabilität des Messinstruments. Der Test sollte daher überarbeitet werden, bevor er erneut eingesetzt wird.
2. Bei der Beobachtung der Unterrichtsreihe zeigte sich, dass insbesondere in der letzten Stunde mehr Zeit für die Diskussion der Aufgaben erforderlich gewesen wäre. So konnten Rückfragen der Schüler nicht in der gewünschten Tiefe behandelt werden und die Zeit auf individuelle Lernschwierigkeiten einzugehen, musste zu Gunsten der Ergebnissicherung stark begrenzt werden.
3. Eine weitere Einschränkung der Interpretierbarkeit der Ergebnisse des Leistungstests ergibt sich aus den Verletzungen der Voraussetzungen: so waren die Daten nicht normalverteilt und auch die Homogenität der Varianzen war nicht gegeben. Die Verwendung der ANCOVA erweist sich gegenüber der Verletzung der Voraussetzungen jedoch als robust, wenn die verglichenen Gruppen annähernd gleich groß sind und den kritischen Wert von $N = 10$ pro Gruppe nicht unterschreiten.
4. Es liegen keine Daten zur längerfristigen Entwicklung vor. So kann z.B. keine Aussage darüber gemacht werden, ob der Lernzuwachs beider Gruppen im Leistungs- und im Konzepttest mittelfristig und längerfristig erhalten bleibt und ob der höhere Lernzuwachs der Treatmentgruppe im Konzepttest mittel- und langfristig stabil ist.
5. Nicht zuletzt handelt es sich um eine recht kleine Stichprobe, die aus einer Schule und einem Schultyp stammt. Die Ergebnisse lassen sich somit nicht auf den Lernerfolg von Schülern an Realschulen plus und Gesamtschulen generalisieren. Die Teststärke ($1-\beta$), also die Wahrscheinlichkeit die H_0 zu widerlegen, wenn sie tatsächlich falsch ist und die H_1 zutrifft, ist für geringe Effekte bei kleinen Stichproben eingeschränkt: So ist die Schätzung der Effektstärke für kleine Stichproben mit großer Unsicherheit behaftet (große Konfidenzintervalle). In der Konsequenz lassen sich für die Ergebnisse des

Leistungstests, der zudem auch nur über eine eingeschränkte Realiblität verfügt, keine zuverlässigen Schlüsse zur Wirkung des Treatments ziehen.

Für den Konzepttest sind die Ergebnisse eindeutiger. Der selbst in der kleinen Stichprobe nachweisbare gefundene starke Effekt ($\omega^2 = .16$) für den höheren Lernzuwachs der Treatmentgruppe weist auf die positive Wirkung der gezielten Thematisierung von wissenschaftlichen Konzepten hin. Schüler, welche hierzu ein gezieltes Treatment erhielten, entschieden sich signifikant häufiger für die korrekten Konzepte und wählten seltener Alternativen, welche weitverbreite Schülervorstellungen beschreiben.

In der Folge wurden nachstehende Änderungen bei der Durchführung der Hauptstudie vorgenommen:

- Der Leistungstest zur Erfassung von Wissen und Problemlösen bei repräsentationsbezogenen Aufgaben wurde grundlegend überarbeitet. Auch der Konzepttest wurde optimiert.
- Um den längerfristigen Lernerfolg erheben zu können, wurde das konzeptuelle Verständnis und die Physikleistung bei repräsentationsbezogenen Aufgaben zu einem dritten Messzeitpunkt erfasst. Aus organisatorischen Gründen wurde ein Zeitabstand von acht Wochen gewählt: acht Wochen betrug der maximale zeitliche Abstand, der es noch ermöglichte, die Studie im Schuljahr 2010/2011 durchzuführen.
- Der Stichprobenumfang wurde deutlich erhöht: So nahmen an der Hauptstudie 21 Klassen aus 10 Schulen teil. Zusätzlich zu acht Gymnasien konnten auch zwei Gesamtschulen für die Teilnahme gewonnen werden.
- Angesichts des klar feststellbaren Effekts des Treatments der Pilotstudie auf die Förderung des konzeptuellen Verständnisses wurde im Anschluss an dieses Ergebnis das Treatment weiterentwickelt. Da sich die entwickelten Aufgaben, welche darauf zielen, weitverbreitete Schülervorstellungen durch die Auseinandersetzung mit fachbezogenen Repräsentationen zu überwinden, offenbar dazu eignen das konzeptuelle Verständnis zu fördern, wurde nun die Art der Aufgabengestaltung in den Blick genommen. Zur Untersuchung dieser Frage wurden zwei Versuchsbedingungen gewählt, die sich im Ausmaß der kognitiven Aktivierung im Umgang mit (multiplen) Repräsentationen unterscheiden. Im Gegensatz zur Pilotstudie, in der sich beide Bedingungen darin unterscheiden, dass Schülervorstellungen thematisiert wurden oder nicht, wurde in der Hauptstudie die Art und Weise mit Repräsentationen umzugehen und das Ausmaß

der kognitiven Aktivierung variiert. Ziel der Studie war es, tiefergehend zu analysieren, wie der Umgang mit Repräsentationen gestaltet werden kann, um das konzeptuelle Verständnis und die Leistung im Umgang mit Repräsentationen zu fördern, anstelle zu prüfen, ob dies überhaupt lernwirksam ist.

- Vor diesem Hintergrund wurde die Unterrichtsreihe von drei auf sechs Unterrichtsstunden erweitert. Die höhere Lernzeit zielte erstens auf einen nachhaltigeren Aufbau von Kompetenzen als dies in der Hälfte der Zeit möglich wäre. Zweitens wurde der Beobachtung der Zeitknappheit in der dritten Unterrichtsstunde in der Pilotstudie Rechnung getragen und drittens wurden zusätzliche Übungszeiten eingeplant, die sich aus der Weiterentwicklung des Treatments ergaben.

2.3 Hauptstudie

2.3.1 Forschungsfragen und Hypothesen

1. Fördern Aufgaben, welche Schüler kognitiv aktivieren, kohärente interne Repräsentationen (mentale Modelle) und externe Repräsentationen zu bilden sowie mit diesen zu operieren, Wissen und Problemlösen in Strahlenoptik (und zwar unmittelbar nach dem Unterricht und mittelfristig)?
2. Fördern Aufgaben, welche Schüler kognitiv aktivieren, kohärente interne Repräsentationen (mentale Modelle) und externe Repräsentationen zu bilden sowie mit diesen zu operieren, unmittelbar nach dem Unterricht und mittelfristig das konzeptuelle Grundverständnis in Strahlenoptik?
3. Steigert der Einsatz der Predict-Observe-Explain-Strategie unter der Perspektive des Lernens mit multiplen Repräsentationen das konzeptuelle Verständnis der Schüler?
4. Führen Aufgaben, welche Schüler kognitiv aktivieren, kohärente interne Repräsentationen (mentale Modelle) und externe Repräsentationen zu bilden sowie mit diesen zu operieren, unmittelbar nach dem Unterricht und mittelfristig zu einer Veränderung der Motivation im Vergleich zu einer Kontrollgruppe?
5. Wirkt sich die Thematisierung von Schülervorstellungen positiv auf die Entwicklung des konzeptuellen Verständnisses aus (im Vergleich zu Physikunterricht zum gleichen Lerninhalt ohne Thematisierung von Schülervorstellungen). Oder anders gefragt: Lassen sich die Ergebnisse der Pilotstudie unter leicht veränderten Bedingungen replizieren?

Hypothese 1 Die Treatmentgruppe erreicht im Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen im Bereich der Strahlenoptik unmittelbar nach der Intervention und ca. zwei Monate später einen höheren Lernzuwachs als die Kontrollgruppe.

Hypothese 2: Die Treatmentgruppe erreicht im konzeptuellen Verständnis im Bereich der Strahlenoptik unmittelbar nach der Intervention und ca. zwei Monate später einen höheren Lernzuwachs als die Kontrollgruppe.

Für die Untersuchung der beiden ersten Hypothesen wurden zwei Versuchsbedingungen gewählt, die sich im Ausmaß der kognitiven Aktivierung im Umgang mit (multiplen) Repräsentationen unterscheiden.

Die Treatmentbedingung beinhaltete kognitiv aktivierende Aufgabenstellungen, welche die Überwindung von gängigen Schülervorstellungen erfordern. In der Kontrollbedingung wurden die gleichen Schülervorstellungen thematisiert und prinzipiell auch vergleichbare Repräsentationsformen verwendet. Schüler in der Kontrollbedingung wurden jedoch nicht gezielt kognitiv aktiviert, eigene Repräsentationen zu erstellen, zu reflektieren und mit gegebenen und selbst erstellten Repräsentationen zu operieren.

Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen im Bereich der Strahlenoptik wurde in einem eigens entwickelten inhaltsspezifischen Leistungstest zu drei Messzeitpunkten erfasst: zum Messzeitpunkt prä (vor der Intervention), zum Messzeitpunkt post (unmittelbar nach der Intervention) und zum Messzeitpunkt follow-up (ca. zwei Monate später).

Das konzeptuelle Verständnis im Bereich der Strahlenoptik wurde ebenfalls in einem eigens entwickelten Konzepttest zu den drei genannten Messzeitpunkten erhoben.

Hypothese 3: Ein Teil der Treatmentgruppe, die ein variiertes Treatment erhielt, welches die Predict-Observe-Explain-Strategie (POE-Sequenz) unter der Perspektive des Lernens mit multiplen Repräsentationen beinhaltet, erreicht im Vergleich zu dem regulären Treatment und im Vergleich zur Kontrollgruppe einen höheren Lernzuwachs im konzeptuellen Verständnis und unmittelbar nach der Intervention und ca. zwei Monate später.

Hypothese 4: Die Entwicklung der Motivation der Schüler in der Treatmentgruppe unterscheidet sich von der Entwicklung in der Kontrollgruppe.

Diese Hypothese wurde als ungerichtete Hypothese formuliert: einerseits könnte sich die intendierte höhere kognitive Aktivierung in der Treatmentgruppe positiv auf die Motivation der Schüler auswirken, andererseits könnte die höhere kognitive Aktivierung in der Treatmentgruppe auch zu erhöhten Belastungen und Anforderungen führen, was die Motivation mindert. Insgesamt steht die

Motivation nicht im Mittelpunkt dieser Studie, da das Treatment vorwiegend auf die Optimierung der kognitiven Lernbedingungen zielt. Als wichtiger Faktor für Lernprozesse wurde sie jedoch mit untersucht.

Hypothese 5: Schüler, die ein Treatment zur Förderung des konzeptuellen Verständnisses erhielten (Treatment- und Kontrollgruppe dieser Studie) erzielten einen höheren Lernzuwachs in Bezug auf das konzeptuelle Verständnis als Schüler, die ein Treatment zur Förderung der Kohärenz (Scheid, 2013) erhielten bzw. der Kontrollgruppe der entsprechenden Studie angehörten.

Zur Beantwortung der fünften Forschungsfrage wurde innerhalb des Projekts die Studie von Scheid (2013) mit den hier berichteten Daten verglichen. Unterschiede zwischen Treatment- und Kontrollgruppe innerhalb der Einzelstudien werden in die Analyse einbezogen, sofern signifikante Unterschiede vorlagen.

Des Weiteren wurden im Rahmen des oben dargelegten Forschungsprogramms (Fragen 1-5) die beiden folgenden Forschungsziele verfolgt.[18]

1. Vor dem Hintergrund, dass für die Domäne der Strahlenoptik im Gegensatz zu anderen Bereichen der Physik, wie der Mechanik (vgl. Hestenes et al., 1992) bislang kein Konzepttest vorliegt, sollte eine tragfähige Fassung eines Konzepttests in der Strahlenoptik erstellt werden. Damit verbundene offene Forschungsfragen betreffen vor allem die Beurteilung, inwiefern die entwickelte Testfassung die Gütekriterien der Objektivität, der Reliabilität und der Validität erfüllt.
2. Im Rahmen der Exploration der Lernprozesse, während der Intervention, wurde der Frage nachgegangen, ob der erfolgreiche Umgang mit zentralen Repräsentationen ein geeigneter Prädiktor für den Lernerfolg unmittelbar nach dem Unterricht darstellt.

Zur Umsetzung der ersten Zielsetzung wurde neben einer ausführlichen Analyse der Itemkennwerte, der Test auf Rasch-Skalierbarkeit geprüft, Korrelationen mit relevanten Außenkriterien (z.B. Physik- und Mathematiknote, figural-räumliches Schlussfolgern etc.) berechnet (Analyse der Kriteriumsvalidität) sowie eine Analyse der Dimensionalität des Tests mittels einer faktorenalytischen Kreuzvalidierung vorgenommen (Konstruktvalidität).

Zur Umsetzung der zweiten Zielsetzung wurde eine per Zufall ausgewählte Teilstichprobe von Arbeitsblättern, die im Rahmen der Intervention eingesammelt wurden, vertiefend analysiert.

18 Fragen, welche mit den genannten Forschungszielen einhergehen, werden im Rahmen dieser Arbeit exploriert, jedoch nicht vollumfänglich beantwortet.

2.3.2 Stichprobe und Design

Die Stichprobe bestand aus $N = 525$ Schülern aus $N = 21$ Schulklassen von 10 Schulen, die von $N = 10$ unterschiedlichen Lehrern unterrichtet wurden. Von den 525 Schülern gingen die Daten von mindestens $N = 443$ Schülern in die Analysen ein. Die übrigen 82 Schüler fehlten an den Tagen, an denen die abhängigen Variablen oder die kognitiven Fähigkeiten erfasst wurden oder es lagen aus diversen Gründen (Schulwechsel zum zweiten Halbjahr) keine vollständigen Informationen zu den Fachnoten vor. In späteren Modellen weicht die zugrundeliegende Datenbasis von $N = 443$ nach oben ab, wenn zu den einzelnen Messzeitpunkten mehr Daten zur Verfügung standen oder aus inhaltlichen Gründen weniger Kovariaten in das Modell eingingen, was entsprechend die Chance erhöht, dass weniger Daten fehlen. Die 525 Schüler waren im Alter von 12 bis 15 Jahren ($M = 13.60$, $SD = 8$ [Monate]) und besuchten entweder ein Gymnasium oder eine Integrierte Gesamtschule (IGS) im Bundesland Rheinland-Pfalz. Zur Stichprobe zählten dabei 17 Gymnasialklassen, $n = 433$ Schüler ($\approx 82\%$ der Gesamtstichprobe) und 4 Klassen an Gesamtschulen, $n = 92$ Schüler ($\approx 18\ \%$ der Gesamtstichprobe). Der Stichprobe gehörten $n = 256$ Jungen und $n = 269$ Mädchen an (vgl. Tabellen Tabelle 3 und 4). Unter den 10 teilnehmenden Physik-Lehrern waren drei Frauen und sieben Männer, die Berufserfahrung der Lehrkräfte variierte zwischen 2 und 30 Jahren. Detaillierte Informationen zur Stichprobe der Lehrer und Schüler können Tabelle 6 im Anhang C9 (auf der Produktseite zu diesem Buch unter www.springer.com), entnommen werden.

Die Umsetzung der Studie war im Vorfeld durch den Datenschutzbeauftragten und die Aufsichts- und Dienstleistungsdirektion des Landes Rheinland-Pfalz geprüft und genehmigt worden. Die Teilnahme an der Studie war freiwillig. Alle Schülerleistungen wurden anonym erfasst. Jeder Schüler erhielt eine Teilnehmernummer, welche er oder sie durchgehend in der Studie verwendete, um die Zuordnung der Testleistungen über unterschiedliche Erhebungsinstrumente und Messzeitpunkte anonym, aber dennoch individuell, auswerten zu können. Die Zuordnung der Teilnehmernummer zu den Namen erfolgte durch die Lehrer und ist der Versuchsleiterin und den beteiligten Mitgliedern der Arbeitsgruppe nicht bekannt. Voraussetzung für den Start der Studie in der jeweiligen Schule war die schriftliche Erlaubnis der Eltern. Nichtteilnehmende Schüler beteiligten sich am Unterricht, ihre Testleistungen wurden jedoch nicht erfasst.

Bei der Hauptstudie handelte es sich um ein einfaktorielles quasi-experimentelles Design. Die Leistung der Schüler wurde zu drei Messzeitpunkten erhoben: unmittelbar vor Beginn der Unterrichtsreihe, direkt nach der Unterrichtsreihe sowie zwei Monate später.

Tabelle 3 und *Tabelle 4* Informationen zur Stichprobe nach Bedingung (Treatment-versus Kontrollgruppe) Klassenstufe, Geschlecht und Schultyp

		Bedingung		
		KG[a]	TG[b]	
Geschlecht	m	121	135	256
	w	129	140	269
		250	275	525

[a]Kontrollgruppe, [b]Treatmentgruppe

		Klassenstufe		
		7	8	
Schulart	[a]GY	171	262	433
	[b]IGS	51	41	92
		222	303	525

[a]Gymnasium, [b]IGS

Auch diese Studie wurde in den regulären Schulalltag eingebunden, was die externe Validität durch die realitätsnahen Bedingungen erhöht. Um den Stundenplan einzuhalten, wurden die Schüler im Klassenverbund als Gruppe den Bedingungen zugeteilt.

An den 10 teilnehmenden Schulen unterrichtete jeweils eine Lehrkraft eine Treatment- und eine Kontrollgruppe derselben Jahrgangsstufe (Parallelklassen), eine Lehrkraft an einem Gymnasium unterrichtete drei Klassen der siebten Jahrgangsstufe, an dieser Schule wurden zwei Klassen der Treatment und eine Klasse der Kontrollbedingung zugeordnet.

Dadurch dass jede Lehrkraft jeweils sowohl in der Treatment- als auch in der Kontrollbedingung unterrichtete, sollte der Faktor der Lehrerpersönlichkeit soweit wie möglich kontrolliert werden.

Auf Basis der Vornoten (Mathematik, Deutsch und Physik) wurden die Klassen abwechselnd der Treatment- und der Kontrollgruppe zugeordnet, um insgesamt möglichst vergleichbare Stichproben zu erhalten. Durch dieses Verfahren kann nur eine Parallelisierung auf Gruppenebene erzielt werden und nicht auf Individualebene. Merkmalsunterschiede zwischen Treatment- und Kontrollbedingung, die von vornherein bestanden, können somit nicht vollständig ausgeschlossen werden.

Zur Prüfung der dritten Hypothese wurde mit einem Teil der Stichprobe (6 Klassen)[19] ein variiertes Treatment durchgeführt, in welchem in der ersten und in der vierten Unterrichtsstunde POE-Sequenzen (White & Gunstone, 1992; Palmer, 1995; Kearney, Treagust, Yeo & Zadnik, 2001; Crouch et al., 2004) eingebaut waren (vgl. Tabelle 5). Die Implementierung der Treatmentvariation zielte darauf ab, zu

19 Diese Teilstichprobe enthielt zwei 8. Klassen eines Gymnasiums, zwei 7. Klassen eines Gymnasiums und zwei 7. Klassen einer Integrierten Gesamtschule.

untersuchen, ob das Verständnis der Bildentstehung vertieft werden kann, wenn die Schüler mit Hilfe der POE-Sequenz kognitiv dazu aktiviert werden, ihre eigenen Vorhersagen, welche sie auch mittels schematischer Repräsentationen darstellen sollten, mit den beobachteten Ergebnissen des Versuchs und der repräsentationalen Deutung der beobachteten Ergebnisse zu kontrastieren. In Parallelklassen, in denen das variierte Treatment umgesetzt wurde, erhielt die Kontrollgruppe den gleichen Unterricht wie die übrigen Kontrollgruppen an den anderen Schulen.

Tabelle 5 Informationen zur Stichprobe je Bedingung und Treatmentform

		Bedingung		
		KG	TG	
Treatmentform	regulär	179	192	371
	variiert	71	83	154
		250	275	525

Beide Klassen, Treatment- und Kontrollgruppe, wurden jeweils von der gleichen Lehrkraft im gleichen Zeitumfang unterrichtet. Die Unterrichtsreihe war ursprünglich auf sechs Schulstunden ausgerichtet. Bei der Umsetzung zeigte sich jedoch, dass alle Lehrer durch unvorhergesehene Unterbrechungen im Schulalltag etwas mehr Zeit benötigten, so dass der Lernstoff auf sieben Unterrichtsstunden (315 min.) verteilt wurde. Dabei wurde die absolute Lernzeit für Treatment- und Kontrollgruppe parallel gehalten. Die Prä- und Posttests fanden jeweils in der Physikstunde vor und nach der Unterrichtseinheit statt und erforderten jeweils 45 min. Die Follow-up Erhebung zwei Monate später erforderte ebenfalls eine Unterrichtsstunde. Zusätzlich wurde noch eine Unterrichts- oder Vertretungsstunde für die Erhebung der kognitiven Fähigkeiten der Schüler aufgewandt.

Die Umsetzung der Studie fand Ende des Schuljahres 2010/2011 statt. Zu diesem Zeitpunkt wurden die Schüler entweder seit August 2010 oder seit Januar 2011[20] erstmals im Fach Physik unterrichtet. Die Unterrichtsreihe wurde begonnen, sobald das Thema der Bildentstehung bei der Sammellinse auch im regulären Curriculum vorgesehen war. Entsprechend hatten bereits alle Schüler einige Vor-

20 Manche Schulen starteten mit dem Physikunterricht erst zum zweiten Halbjahr (Epochalunterricht).

kenntnisse zu den Konzepten der geradlinigen Lichtausbreitung, zur Streuung, zur Brechung an Grenzflächen und zur physikalischen Sehvorstellung erworben.

2.3.3 Operationalisierung des Treatments

2.3.3.1 Genereller Überblick

Für die Intervention wurden zwei Versuchsbedingungen gewählt, die sich im Ausmaß der kognitiven Aktivierung im Umgang mit (multiplen) Repräsentationen unterscheiden. Die Unterrichtsreihe umfasste in der Planung sechs Schulstunden:

1. eine Einführungsstunde zum Thema Brechungsvorgänge an der Sammellinse. In dieser Stunde zeigte die jeweilige Lehrkraft den Schülern ein Demonstrationsexperiment, bei dem parallel einfallende Lichtstrahlen auf eine Sammellinse treffen und eine optische Tafel streifen; auf Basis des Demonstrationsexperiments befassten sich die Schüler mit den Grundlagen der Konstruktion von Strahlengängen,
2. der Durchführung des Schülerexperiments zur Entstehung reeller Bilder bei der Sammellinse inklusive der Auswertung des Experiments in der zweiten Stunde,
3. die Konstruktion der unterschiedlichen Bildfälle (vergrößertes, verkleinertes und gleich großes Bild) in der dritten Stunde,
4. die Thematisierung und Konstruktion des virtuellen Bildes in der vierten Stunde,
5. der Einführung des Abbildungsgesetzes mit verbaler und geometrischer Veranschaulichung der Zusammenhänge in der fünften Stunde und
6. der Abschlussstunde, in der die Bildkonstruktion und das Abbildungsgesetz vertiefend geübt wurden.

Kognitiv aktivierende Lernstrategien zielen darauf, die Lernenden durch anspruchsvolle Aufgaben dazu anzuregen, die erlernten physikalischen Vorgänge mental durchzuspielen, also geeignete mentale Modelle aufzubauen. Der Aufbau solcher adäquaten mentaler Modelle sollte, sofern dieser Schritt gelingt, zum einen dazu führen, Schülervorstellungen zu überwinden, da Inkonsistenzen mit einer adäquaten Modellvorstellung inkompatibel sind (vgl. Gentner & Gentner, 1983; diSessa, 1983, 1988, 1993; Botzer & Reiner, 2005; Mortimer & Buty, 2009) und sich zum anderen auch in externen Repräsentationen niederschlagen (vgl. Cox, 1999), die wiederum dazu genutzt werden können, Problemlöseaufgaben zu bearbeiten (vgl. Gentner & Gentner, 1983; Ohlsson, 1992; Zhang, 1997).

Tabelle 6 Überblick über Formen der kognitiven Aktivierung in der Treatmentbedingung

Art der kognitiven Aktivierung in der TG	Theoretischer / empirischer Bezug:	Aufgaben[a] siehe Anhang C1[b]	Std.
Kognitive Aktivierung durch das Erstellen eigener Repräsentationen	Eigene Gedanken, Konzepte, Ideen und Lösungswege erläutern (Hiebert & Wearne, 1993; Shayer & Adhami 2007).	Aufgabenblatt 1 [Mat2_K1] A1 Experimentieranleitung [Mat5_K1] Aufgabenblatt 4 [Mat7_K1] Aufgabenblatt 5 [Mat8_K1] Aufgabenblatt 7 [Mat10_K1] A2 und A3 Aufgabenblatt 8 [Mat12_K1]	1 2 3 4 5 5
Abgleich der eigenen Vorstellung mit der Lösung. Der Abgleich mit der Lösung bezweckte, die eigene Sichtweise zu reflektieren und die eigenen Verarbeitungsschritte entsprechend anzupassen	Beziehungen zwischen Vorwissen bzw. dem derzeitigem Lernstand und neuen Lerninhalten herstellen (Hiebert & Wearne, 1993; Baumert & Kunter, 2011).	Experimentieranleitung [Mat5_K1] Aufgabenblatt 4 [Mat7_K1] Aufgabenblatt 7 [Mat10_K1] A2 und A3	2 3 5
Kognitive Aktivierung durch die Anforderung, Repräsentationen aufeinander zu beziehen bzw. ineinander zu übersetzen, die in unterschiedlichen Formaten vorliegen oder auf unterschiedlichen Abstraktionsebenen liegen	Beziehungen zwischen verschiedenen Repräsentationsformen herstellen (Hiebert & Wearne, 1993; Stein & Lane 1996).	Aufgabenblatt 3 [Mat6_K1] Aufgabenblatt 4 [Mat7_K1] Aufgabenblatt 6 [Mat9_K1] Aufgabenblatt 7 [Mat10_K1], A2 und A3 Aufgabenblatt 8 [Mat12_K1]	2 3 4 5 5
Repräsentationales Fading-out	Beziehungen zwischen dem derzeitigem Lernstand und neuen Lerninhalten herstellen (Hiebert & Wearne, 1993; Baumert & Kunter, 2011).	Aufgabenblatt 4 [Mat7_K1] (Konstruktion der Bildfälle) Aufgabenblatt 7: [Mat10_K1], A2 und A3	3 5
Förderung der kognitiven Flexibilität im Umgang mit Repräsentationen	Beziehungen zwischen verschiedenen Repräsentationsformen herstellen (Hiebert & Wearne, 1993; Stein & Lane 1996).	Aufgabenblatt 9 [Mat13_K1], A1 und A2	6
Einsatz von Aufgaben, die darauf angelegt sind, ein internes mentales Modell physikalischer Abläufe oder	Beziehungen zwischen verschiedenen Repräsentationsformen herstellen (Hiebert & Wearne, 1993; Stein & Lane 1996).	Aufgabenblatt 3 [Mat6_K1] Aufgabenblatt 6 [Mat9_K1]	2 4 5

Zusammenhänge auf Basis depiktionaler schematischer Repräsentationen zu bilden, welche die Schüler die physikalischen Abläufe bzw. Zusammenhänge in kohärenter Weise intern dynamisch simulieren lassen.		Aufgabenblatt 7 [Mat10_K1], A2 und A3	
Treatmentvariation			
POE-Sequenz, Vorhersage unter Verwendung von Repräsentationen in unterschiedlichen Formaten, sowie Abgleich der eigenen Vorstellung mit der Lösung.	Kognitive Konflikte initiieren (Baumert & Kunter, 2011); Beziehungen zwischen verschiedenen Repräsentationsformen herstellen (Hiebert & Wearne, 1993; Stein & Lane 1996); einen prozessorientieren Umgang mit Problemen fördern, Klieme et al., 2006)	Demonstrationsexperiment an der optischen Scheibe [Zusatz-1-K1] TG und [Zusatz-2-K1] TG	14

[a]Mehrfachnennungen möglich
[b]auf der Produktseite zu diesem Buch unter www.springer.com

Die Art der im Unterricht verwendeten Repräsentationen und die angesprochenen Schülervorstellungen sowie der Einsatz eines Demonstrations- und eines Schülerexperiments waren in beiden Versuchsbedingungen konstant gehalten. Folgende Schülervorstellungen wurden sowohl in der Treatment- als auch in der Kontrollgruppe behandelt:

Tabelle 7 Überblick über thematisierte Konzepte und damit verbundene Schülervorstellungen in der Treatment- (TG) und der Kontrollbedingung (KG)

Thematisierte Grundkonzepte	Adressierte Schülervorstellungen (TG und KG)	Aufgaben[a] siehe Anhang C1[b]	Std.
Wissenschaftstheoretisch adäquates Lichtverständnis; Physikalische Sehvorstellung und geradlinige Lichtausbreitung	Verwechslung Lichtbündel und Lichtstrahl Verständnisprobleme mit der geradlinigen Lichtausbreitung; Fokussierung auf ausgezeichnete Strahlen	Aufgabenblatt 1 [Mat2_K1] A1 Aufgabenblatt 3 [Mat6_K1] TG	1 2
Entstehung reeller Bilder bei der Sammellinse, Konzept der Punk-zu-Punkt-Abbildung	Fehlerhaftes Verständnis der Funktionsweise von Sammellinsen, fehlerhafte bildliche Repräsentation, Vorbeugung einer holistischen Konzeption des Abbildungsvorgangs durch Einüben der Strahlenkonstruktion	Aufgabenblatt 3 [Mat6_K1] TG Aufgabenblatt 4 [Mat7_K1] (Konstruktion der Bildfälle)	2 3
Virtuelle Bilder bei der Sammellinse	Probleme beim Verständnis der Entstehung (und Lage) virtueller Bilder	Aufgabenblatt 5 [Mat8_K1]	4
Entstehung reeller Bilder bei der Sammellinse	Bei Abdeckung der oberen bzw. unteren Hälfte der Linse wird entsprechend die obere bzw. untere Hälfte des Bildes angeschnitten	Aufgabenblatt 7: [Mat10_K1], A2 (Hinführung) und A3	5
Entstehung reeller Bilder bei der Sammellinse: Aspekt Funktion des Schirms / Ort des Bildes	Schwierigkeiten beim Verständnis des Bildortes: Vorstellung, das Bild sei auf der Linse; nur wenige Lernende erkennen, dass sich das Bild ohne Schirm an der gleichen Position befindet wie der Schirm.	Aufgabenblatt 9 [Mat13_K1], A1	6
Treatmentvariation			
Entstehung reeller Bilder bei der Sammellinse	Bei Abdeckung der oberen bzw. unteren Hälfte der Linse wird entsprechend die obere bzw. untere Hälfte des Bildes angeschnitten	[Zusatz-1-K1] TG und [Zusatz-2-K1] TG (Demonstrationsversuch „abgedeckte Sammellinse“)	4

[a] Mehrfachnennungen möglich
[b] auf der Produktseite zu diesem Buch unter www.springer.com

Solche kognitiven Lernstrategien beinhalten auch eine metakognitive Komponente (vgl. Schrader & Helmke 2006, S. 638 f.). Durch die Aufforderung, den physikalischen Sachverhalt zeichnerisch und schriftlich darzustellen, verlangen sie von den Lernenden, ihre Vorstellungen zu reflektieren sowie mit den erstellten Repräsentationen kognitiv zu operieren. Die Operation mit den erstellten externen Repräsentationen beinhaltet, deren Zweckmäßigkeit zur Lösung von physikalischen Problemen zu prüfen, und zielt somit darauf, die eigenen Verarbeitungsschritte anzupassen.

Das vollständige Unterrichtsmaterial ist in Anhang C1, die zugehörige Unterrichtsplanung in Anhang C2 (auf der Produktseite zu diesem Buch unter www.springer.com) zu finden. Zusammenfassungen zur Operationalisierung finden sich in Tabelle 6 und Tabelle 7. Zur guten Lesbarkeit der Aufgabentexte sei ebenfalls auf Anhang C1 unter www.springer.com hingewiesen.

2.3.3.2 Operationalisierung des Treatments in der ersten Unterrichtsstunde

Unterschiede in der kognitiven Aktivierung im Umgang mit Repräsentationen bezogen sich in der ersten Stunde auf eine unterschiedliche Thematisierung der Begriffe „Lichtbündel“ und „Lichtstrahl“. Beiden Gruppen wurde folgendes Demonstrationsexperiment gezeigt: auf einer optischen Tafel ist eine Sammellinse und eine Lampe mit einer Schlitzblende montiert. Durch die Schlitzblende werden schmale parallel einfallende Lichtbündel erzeugt, welche die optische Tafel streifen (die Lichtbündel sind durch die Streuung an der optischen Tafel zu sehen). Die Sammellinse vereinigt das achsenparallel einfallende Licht im Brennpunkt. Nachdem sich die parallel einfallenden Lichtbündel im Brennpunkt getroffen haben, divergieren sie. Bezogen auf das Demonstrationsexperiment sollten sich die Schüler den Unterschied zwischen Lichtbündel und Lichtstrahl klar machen.

Während man im Demonstrationsexperiment Lichtbündel sehen kann, die eine optische Tafel streifen, handelt es sich bei Lichtstrahlen um ein gedankliches Konstrukt, das zur Beschreibung und zeichnerischen Darstellung von Lichtbündeln dient. Die hiermit verbundene Schülervorstellung betrifft eine materialistische Konzeption von Licht, in welcher der Begriff (Licht-)„Strahl“ nicht im mathematischen Sinn aufgefasst wird, sondern als konkreter Strom, analog einem Flüssigkeitsstrom verstanden wird (vgl. Andersson & Kärrqvist, 1983; Reiner et al., 2000, S. 14 f.).

Gerade letztere Vorstellung zeigt, dass Lernende mit wenig physikalischer Vorbildung ontologisch grundlegend verschiedene Vorstellungen von Licht besitzen.

In beiden Gruppen wird diese Schülervorstellung durch eine verbale Repräsentation angesprochen, die sich auf das Demonstrationsexperiment bezieht (vgl. Abbildung 20). Durch das offene Frageformat in der Treatmentgruppe wird diese jedoch kognitiv aktiviert, eine eigene verbale Repräsentation zu erstellen und zu prüfen, während die Kontrollgruppe aus zwei gegebenen verbalen Repräsentationen die zutreffende Repräsentation wählen muss.

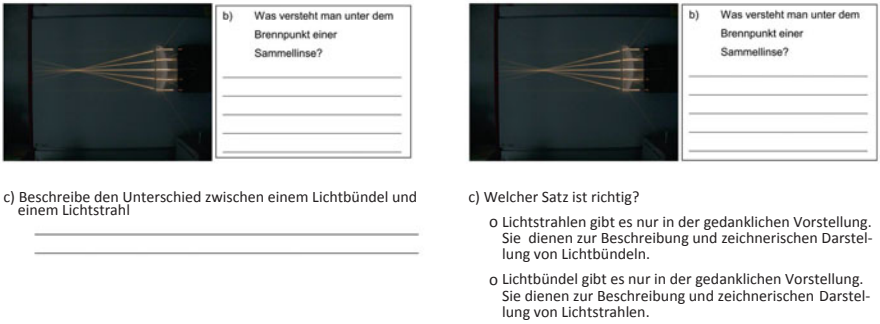


Abbildung 20: Aufgabenblatt 1 - Die Sammellinse[Mat 2_K1] TG bzw. [Mat 2_K2] KG links Aufgabe der Treatmentgruppe (TG), rechts Aufgabe der Kontrollgruppe (KG)

2.3.3.3 Treatmentvariation in der ersten Unterrichtsstunde

In der Treatmentgruppe wurden nun zwei kurze POE-Sequenzen eingebaut. Vor der Vorführung des Demonstrationsexperimentes wurden die Schüler gefragt, wie das Licht verläuft, wenn man die Lampe anschaltet und eine Linse davor setzt. Die Schüler sollten den Verlauf des Lichts vorhersagen und ihre Vermutungen zum Strahlenverlauf in einer schematischen Repräsentation an der Tafel zeichnen (predict).

Anschließend wurde den Schülern das Experiment demonstriert und die Beobachtung der Schüler festgehalten (observe). Abschließend wurde die Funktionsweise der Linse erklärt und die zentralen Begriffe wie „der Brennpunkt *F*“ und „die Brennweite *f*“ sowie der Begriff der optischen Achse erarbeitet (explain). Das Unterrichtsmaterial zur Treatmentvariation kann in Anhang C3 und die Unterrichtsplanung zum variierten Treatment in Anhang C4 jeweils auf der Produktseite zu diesem Buch unter www.springer.com eingesehen werden.

In einer zweiten Sequenz wurde die Auswirkung verschiedener Brennweiten thematisiert. So sollten die Schüler vorhersagen, was passiert, wenn man eine stärker gekrümmte Linse verwendet. Auf Basis der Schülerantworten erstellte die

Lehrkraft entsprechende Tafelbilder (predict). Anschließend führte die Lehrkraft den Schülern das Experiment vor, wobei die Beobachtungen der Schüler festgehalten wurden (observe). Abschließend klärten Schüler und Lehrer im Dialog das Ergebnis (explain).

In der Kontrollgruppe wurde der Unterrichtslauf belassen: Die Schüler erarbeiteten auf Basis des Demonstrationsexperiments die zentralen Begriffe ohne POE-Sequenz, in der die Schüler ihre Vorhersagen verbal und zeichnerisch darstellten und anhand des Experiments nachbereiteten.

2.3.3.4 Operationalisierung des Treatments in der zweiten Unterrichtsstunde

In der zweiten Unterrichtsstunde führten sowohl die Treatment- als auch die Kontrollgruppe ein Schülerexperiment zur Entstehung reeller Bilder bei der Sammellinse durch. Analog zu dem Versuch zur Bildentstehung am Hohlspiegel der Pilotstudie montierten sie zunächst eine Kerze (auf einem Kerzenteller), eine Sammellinse und einen Schirm auf einer optischen Bank. Durch das Verschieben von Kerze, Linse und Schirm konnten die Abstände zwischen Kerze und Linse sowie zwischen Linse und Schirm eingestellt werden. Durch die Wahl geeigneter Abstände war es für die Schüler nun möglich, ein reelles, verkleinertes, vergrößertes oder gleichgroßes Bild der Kerzenflamme auf dem Schirm aufzufangen. Die Aufgabe bestand darin, diese Bilder zu finden und die Flammengröße von Kerze und Bild zu messen.

Da das Experiment die Unterrichtszeit von 45 min. nahezu ausfüllte, konzentrierte sich das Herausarbeiten des Unterschieds zwischen Treatment- und Kontrollgruppe auf das Nacharbeiten in der Hausaufgabe.

Beide Gruppen befassten sich zur Wiederholung mit dem Versuchsaufbau und den hiermit verbundenen physikalischen Begriffen: „Bildweite“, „Gegenstandsweite“, „Bildgröße“ und „Gegenstandsgröße“. Eine Schülervorstellung wurde in diesem Kontext nicht adressiert.

In beiden Gruppen erforderte es die Aufgabe, zwei unterschiedliche Ebenen miteinander zu verknüpfen:

- das reale Experiment (Ebene der konkreten Gegenstände), auf welche die Skizze des Versuchsaufbaus (Repräsentation des Versuchsaufbaus) referiert,
- sowie die Beschriftung des Versuchsaufbaus, bei der die physikalischen Größen (deskriptiv abstrakte Repräsentationen) auf die Ebene des realen Experiments und ihre Repräsentation in der Skizze bezogen werden. In den Unterrichts-

beobachtungen zeigte sich, dass diese Aufgabe für die Schüler keineswegs so trivial war, wie sie erscheinen mag: Probleme ergaben sich v.a. in der Zuordnung von Gegenstandsweite und Bildweite. Hier kam es des Öfteren zu Verwechslungen.

Die kognitive Aktivierung in der Treatmentgruppe bestand darin, die Skizze selbst zu erstellen, während der Kontrollgruppe die Skizze vorgegeben wurde. Das eigene Erstellen der Skizze stellt nicht nur eine kognitive Aktivierung im Sinn eigener Tätigkeit dar. Die kognitive Aktivierung der Treatmentgruppe sollte als Ergebnis das aktive Abrufen der Lerninhalte zeigen. Aus der kognitionspsychologischen Forschung ist bekannt, dass selbstgenerierte Elaborationen die elaborative Verarbeitung und damit die Behaltensleistung der Lernenden fördern (vgl. Stein & Bransford, 1979, zit. n. Anderson, 1996, S. 187). Auf metakognitiver Ebene verlangt die Fragestellung in der Treatmentgruppe von den Schülern, ihr Verständnis auch von der Funktion des Versuchsaufbaus im Abgleich mit der Lösung zu reflektieren (vgl. Abbildung 21).

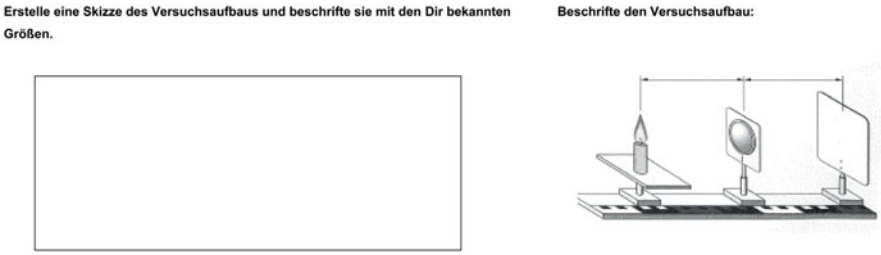


Abbildung 21: [Experimentieranleitung Mat. 5_ K1] bzw. [Mat5 1], Abbildung rechts entnommen aus Backhaus et al. (2008). Fokus Physik Gymnasium Rheinland-Pfalz, S. 35

Die zweite Hausaufgabe bestand in der Bearbeitung eines Arbeitsblatts (Aufgabenblatt 3), das auf die Bildkonstruktion hinführte.

In beiden Gruppen arbeiteten die Schüler mit einer schematischen Repräsentation, in der eine Kerze vor einer Sammellinse dargestellt war. Anhand dieser Repräsentation wurde die physikalische Sehvorstellung thematisiert.

Zum Kern der physikalischen Sehvorstellung zählt, dass beleuchtete Gegenstände durch Streuung Licht abstrahlen (Wiesner, 1994, S. 7) und dieses Licht

ins Auge fällt. Fehlt eine solche physikalische Sehvorstellung, ergeben auch die Strahlenkonstruktionen in der Strahlenoptik wenig oder überhaupt keinen Sinn (Wiesner, 1986, 1992b).

Aus der Fokussierung auf die Konstruktion der ausgezeichneten Strahlen ergibt sich zudem häufig ein im Unterricht selbst erzeugtes Problem. Viele Schüler beachten folgendes nicht: Die ausgezeichneten Strahlen werden zur Bildkonstruktion verwendet; bei dem physikalischen Phänomen der Bildentstehung im konkreten Experiment hingegen trägt alles Licht, das auf die Sammellinse fällt, zur Bildentstehung bei. An dieser Stelle bestehen daher Verständnisschwierigkeiten auf zwei Ebenen:

- der Unterscheidung Lichtbündel versus Lichtstrahl und
- einer adäquaten Vorstellung der Lichtausbreitung.

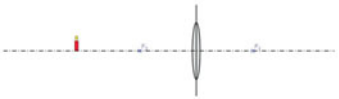
Hieraus können spätere Verständnisschwierigkeiten resultieren, die sich in Schülervorstellungen äußern, wie etwa in der Vorstellung, wenn man eine Blende vor die Linse hält, werde ein Teil des Bildes abgeschnitten (Wiesner, 1992b ,S. 288). Verfügen die Schüler hingegen über ein adäquates Modell der Lichtausbreitung und kennen den Unterschied zwischen einem Lichtbündel und dem idealisierten Modell eines Lichtstrahls, könnte diesen Verständnisschwierigkeiten möglicherweise vorgebeugt werden. Die im Folgenden entwickelte Instruktion sollte auf ein solches adäquates mentales Modell hinführen.

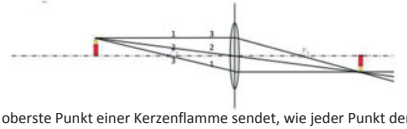
1. Die Treatmentgruppe wurde in einem ersten Schritt aufgefordert, vom obersten Punkt der Kerze Strahlen in all diejenigen Richtungen einzuzeichnen, in denen die Kerze gemäß ihrer Vorstellung Licht abstrahlt.
2. Im zweiten Schritt sollten die Schüler die ausgezeichneten Strahlen, Brennpunktstrahl, Mittelpunktstrahl und Parallelstrahl, einzeichnen. Die Durchführung beider Schritte in der gegebenen Reihenfolge sollte den Schülern helfen, durch eigenes Operieren mit der Repräsentation diesen Zusammenhang zu erkennen.
3. In einem dritten Schritt wurden die Schüler aufgefordert, explizit zeichnerisch zu verdeutlichen, dass auch andere Strahlen außer den ausgezeichneten Strahlen zur Bildentstehung beitragen.

Die Kontrollgruppe erhielt parallel hierzu die gleiche Repräsentation. Anstelle sich ein adäquates Modell der Lichtausbreitung durch das Operieren mit der schematischen Repräsentation der Kerze vor der Sammellinse zu erarbeiten, wurde die Kontrollgruppe direkt durch einen kurzen Text instruiert (vgl. Abbildung 22).

b) Zeichne von dem obersten Punkt der Kerze Strahlen in alle Richtungen in die Abbildung unten ein, von denen du glaubst, dass es Strahlen gibt.

c) Es gibt einige Strahlen, die auf die Sammellinse fallen und deren Verlauf du leicht angeben kannst. Zeichne diese Strahlen in die Abbildung unten ein und benenne sie! Welche Lichtstrahlen tragen ebenfalls zur Bildentstehung bei? Zeichne ein.





Der oberste Punkt einer Kerzenflamme sendet, wie jeder Punkt der Flamme, Licht nach allen Seiten aus.
Ein Teil des Lichts fällt auf die Sammellinse.
Alles Licht, das von der Kerze auf die Sammellinse fällt, trägt zur Bildentstehung bei. Von bestimmten Strahlen kennst du den Verlauf.

b) Benenne dies Strahlen (siehe Abbildung oben)!

1_____
2_____
3_____

Neben achsenparallel einfallenden Lichtbündeln und Lichtbündeln durch den optischen Mittelpunkt, trägt auch alles andere Licht, das auf die Sammellinse trifft, zur Bildentstehung bei.

Abbildung 22: Aufgabenblatt 3: Bildkonstruktion für die Sammellinse [Mat.6 _K1] bzw. [Mat. 6 _ K2] links Aufgaben der TG, rechts Aufgaben der KG 1

Auch in dieser Gruppe wurde den entsprechenden Verständnisschwierigkeiten also vorgebeugt.

2.3.3.5 Operationalisierung des Treatments in der dritten Unterrichtsstunde

Basierend auf der Hinführung auf die Strahlenkonstruktion durch die Hausaufgabe und die Einführung der ausgezeichneten Strahlen in Aufgabenblatt 2, lernten die Schüler in der dritten Stunde unterschiedliche Bildfälle (vergrößerte, verkleinerte und gleich große Bilder) zu konstruieren, die den Beobachtungen im Schülerexperiment entsprachen[21] (siehe Aufgabenblatt 4: Die Konstruktion der Bildfälle [Mat. 7_K1] bzw. [Mat.7_K2] in Anhang 6 auf der Produktseite zu diesem Buch unter www.springer.com).

In der Treatmentgruppe setzten sich die Schüler in einem dreischrittigen Verfahren mit der Konstruktion auseinander. Dabei erhielten die Schüler ein repräsentationales „Fading-out“[22] mit dem Ziel, die Schüler sowohl beim Erstellen

21 Die Schüler verwendeten für die Strahlenkonstruktionen die Hauptebene der Linsen. Der tatsächliche Strahlenverlauf infolge der Brechung an den beiden Außenflächen der Linse wurde aus Zeitgründen nicht thematisiert und ist auch nicht im Lehrplan des Landes Rheinland-Pfalz für das Fach Physik vorgesehen (vgl. Ministerium für Bildung, Wissenschaft, Jugend und Kultur. Rheinland-Pfalz. Lehrplan-Entwürfe Lernbereich Naturwissenschaften Biologie Physik Chemie).

22 Beim Fading-out (Englisch „fade out“ = abblenden, ausblenden) werden anfänglich gegebene Hilfestellungen allmählich verringert (Prinzip abnehmender Lernhilfen), auf diese Weise wird den wachsenden Fähigkeiten der Lernenden Rechnung getragen und eigenständiges Arbeiten ermöglicht und eingefordert (vgl. Reinmann-Rothmeier & Mandl, 1994, S. 632; Schnotz, 2006b, S. 126).

der Konstruktion als auch beim Verknüpfen der Konstruktion (schematische Repräsentation) mit dem Merksatz zu unterstützen. Der Merksatz, der als deskriptiv verbale Repräsentation einzuordnen ist, repräsentiert den Zusammenhang zwischen Bildweite, Gegenstandsweite, Brennweite und Bildgröße in einer verallgemeinernden und abstrakten Form.

1. In einem ersten Schritt erhielten die Schüler eine Repräsentation, die analog zur schematischen Repräsentation in Aufgabenblatt 3 war. Sie wurden aufgefordert, das Bild der dargestellten Kerze vor einer Sammellinse zu konstruieren und zu benennen, um welchen Fall es sich handelt (vergrößertes, verkleinertes oder gleich großes Bild). Zu Zwecken der Datenerhebung wurde das erste Arbeitsblatt nach einer vorgegeben Bearbeitungszeit von 5 min. eingesammelt.
2. In einem zweiten Schritt erhielten die Schüler eine Teillösung der Strahlenkonstruktion und sollten, die gefragten Größen, Gegenstands- und Bildweite sowie Gegenstands- und Bildgröße, abmessen und den Merksatz ergänzen. Die Ergänzung des Merksatzes erforderte von den Schülern, die schematisch-bildhafte Repräsentation der Strahlenkonstruktion im Hinblick auf die allgemeinen Verhältnisse von Bildweite, Gegenstandsweite, Brennweite und Bildgröße zueinander zu abstrahieren und in eine verbale Repräsentation zu übersetzen. Auch das zweite Arbeitsblatt wurde nach 5 Minuten eingesammelt.
3. In einem dritten Schritt bekamen die Schüler zur Ergebnissicherung für ihre persönlichen Unterlagen eine vollständige Lösung der Aufgaben mit einem Lösungsblatt, welches die Konstruktion, die Messergebnisse der Zeichnung und den vollständigen Merksatz enthielt.

Das Fading-out fand zeitlich auf zwei Ebenen statt:

- Zum einen erhielten die Schüler mit der Teillösung eine Rückmeldung, in der sie abgleichen konnten, inwiefern sie das erste Aufgabenblatt korrekt bearbeitet hatten. Schüler, denen die Strahlenkonstruktion Probleme bereitet hatte, erhielten durch die Teillösung sowohl eine Hilfestellung bei der Konstruktion als auch die Chance, die Konstruktion im zweiten Anlauf zu vervollständigen und auf Basis der Lösung die zweite Aufgabe, den Merksatz, zu ergänzen.
- Zum anderen wurden die Teillösungen des zweiten Aufgabenblatts zunehmend unvollständiger sowie Aufgaben und die Ergänzung des Merksatzes von Mal zu Mal offener gestaltet: So war im ersten Durchgang bei dem gleichgroßen Bild der Parallelstrahl vollständig und Brennpunktstrahl sowie Mittelpunktstrahl bis zur Hauptebene (Linsenmitte) konstruiert. Im Merksatz musste

lediglich der Zusammenhang hergestellt werden, dass Gegenstandsweite und Bildweite einander und beide jeweils der doppelten Brennweite entsprechen. Um das Erkennen dieses Zusammenhangs aus der Konstruktion (depiktionale Repräsentation) zu erleichtern, war die doppelte Brennweite eingezeichnet und benannt.

- Im zweiten Durchgang (verkleinertes Bild) waren in der Teillösung - die ausgezeichneten Strahlen - nur bis zur Hauptebene konstruiert. Im Merksatz mussten neben den gefragten Größen auch Texte ergänzt werden.
- Im dritten Durchgang (vergrößertes Bild) waren in der in der Teillösung die ausgezeichneten Strahlen ebenfalls nur bis zur Hauptebene konstruiert. Den Merksatz sollten die Schüler entsprechend der Lösungsmodelle, die sie zuvor erhalten hatten, selbst formulieren.

Die Treatmentgruppe erhielt parallel hierzu einen ausführlichen Lehrervortrag, in dem die Strahlenkonstruktion und die Merksätze erläutert wurden. Die Aufgabe der Schüler bestand darin, die dargestellten Zusammenhänge zu verstehen und die gefragten Größen (Bildweite, Gegenstandsweite, Gegenstandsgröße und Bildgröße) auf den ausgeteilten Arbeitsblättern zu ergänzen.

Prinzipiell arbeiteten also beide Gruppen mit den gleichen Repräsentationen: der Strahlenkonstruktion, der Ergänzung der konkreten Zahlenwerte gefragter physikalischer Größen und dem Merksatz, in dem allgemeinen Zusammenhänge zwischen den involvierten physikalischen Größen abstrakt verbal repräsentiert werden.

Die Treatmentgruppe wurde jedoch kognitiv aktiviert, die entsprechenden Repräsentationen teilweise selbst zu erstellen, Repräsentationen zu ergänzen und eigenständig unterschiedliche Repräsentationen aufeinander zu beziehen bzw. die schematische Repräsentation (über den Zwischenschritt des Abmessens der konkreten Zahlenwerte) in eine abstrakte generalisierende verbale Repräsentation zu überführen.

In der Hausaufgabe wurde die Idee der Vervollständigung eines repräsentationalen „Lückentextes“ aufgegriffen. Während die Treatmentgruppe die Ergebnisse in einer Tabelle über die Bildfälle zusammenfassen sollte, bestand die Aufgabe der Kontrollgruppe darin, die gegebenen Bildfälle mündlich erläutern zu können. Die tabellarische Übersicht war hier als Lernmaterial gedacht.

2.3.3.6 Operationalisierung des Treatments in der vierten Unterrichtsstunde

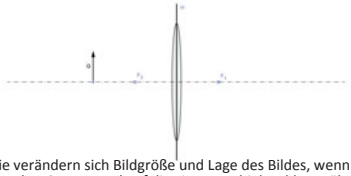
Thema der vierten Unterrichtsstunde war die Konstruktion des virtuellen Bildes. Mit dem virtuellen Bild wurde die Bildentstehung bei der Sammellinse vervollständigt. Die Konstruktion des virtuellen Bildes wurde nicht im Leistungstest abgefragt und spielt daher für die Operationalisierung des Gesamttreatments im Vergleich zu Aufgaben, die auf die Vertiefung der Konstruktion der reellen Bilder zielen, eine untergeordnete Rolle.

Da es sich bei dem Konzept des virtuellen Bildes um einen relativ anspruchsvollen und abstrakten Lerninhalt handelt, der vielen Schülern Lernschwierigkeiten bereitet (vgl. Wiesner, 1986), wurde sowohl die Kontrollgruppe als auch die Treatmentgruppe schrittweise auf die Konstruktion des virtuellen Bildes hingeführt.

Während der Kontrollgruppe die Konstruktion weitgehend vorgegeben war, sollte die Treatmentgruppe die Strahlenkonstruktion in mehreren Schritten bearbeiten und die gegebenen Konstruktionen ergänzen. Aufgabenblatt 5: Das virtuelle Bild [Mat. 8 _ K1] bzw. [Mat. 8 _K2]. Auch hier wurde die Treatmentgruppe kognitiv dazu aktiviert, sich intensiver mit der Konstruktion auseinanderzusetzen. Während die Treatmentgruppe unter Berücksichtigung der selbst erstellen Konstruktionen sich die Begriffe „virtuelles Bild“ im Gegensatz zu einem „reellen Bild“ selbst erarbeiten mussten, wurde der Kontrollgruppe dieser Unterschied vom Lehrer erklärt. Die Erklärung der Unterscheidung forderte dabei, Schlüsse aus der unter Anleitung selbst erstellten Konstruktion zu ziehen und diese in Worte zu fassen (also zwei Repräsentationen, die in unterschiedlichen Formaten vorliegen, aufeinander zu beziehen). In der Kontrollgruppe wurden diese Schritte vom Lehrer übernommen. Zum Abschluss des Themas „virtuelle Bilder“ bearbeiteten beide Gruppen eine Aufgabe zu einer schematischen Abbildung des Versuchsaufbaus des virtuellen Bildes. Während die Treatmentgruppe gebeten wurde, einen Versuchsaufbau zum virtuellen Bild mit eigenen Worten zu erklären und die vorliegende Versuchsskizze zu erläutern, sollte die Kontrollgruppe die Versuchsskizze lediglich beschriften.

Nachdem nun alle Fälle der Bildentstehung vollständig thematisiert waren und auch in der tabellarischen Übersicht vervollständigt werden konnten, wurde die übrige Unterrichtszeit genutzt, um die Konstruktion des reellen Bildes zu vertiefen. Die Treatmentgruppe bearbeitete eine Aufgabe, mit dem Ziel, ein internes mentales Modell der Bildentstehung auf Basis der Strahlenkonstruktion zu bilden, das den Abbildungsvorgang in kohärenter Weise intern dynamisch simuliert.

A1
Ein Gegenstand - dargestellt durch einen Pfeil - steht in einem Abstand *g = 2f* vor einer Sammellinse



a) Wie verändern sich Bildgröße und Lage des Bildes, wenn man den Gegenstand auf die Linse zuschiebt, d.h. *g* nähert sich *f*? Gib die ungefähre Lage und Größe des Bildes in der Abbildung an (ohne Konstruktion) und kennzeichne den Vorgang durch Pfeile.

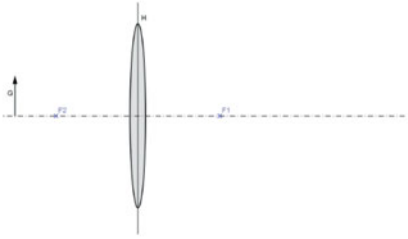
b) Welche Abbildungseigenschaften liegen vor, wenn *g* innerhalb der einfachen Brennweite *f* liegt (siehe Tabelle)?

c) Wie verändern sich Bildgröße und Lage des Bildes, wenn man den Gegenstand weiter von der Linse weg schiebt *(g > 2f)*? Gib die ungefähre Lage und Größe des Bildes in der Abbildung an (ohne Konstruktion) und kennzeichne den Vorgang durch Pfeile.

d) Was passiert, wenn man den Gegenstand sehr weit von der Linse entfernt? Welchem Wert nähert sich die Bildweite?

A1
a) 6 cm vor einer Sammellinse steht ein Gegenstand mit der Größe von 2 cm. Die Linse hat eine Brennweite von 4 cm. Bestimme durch die Zeichnung Ort und Größe des Bildes (Abb. 2).

b) Nadine möchte den Brennpunkt ihrer Lupe (Sammellinse) bestimmen.



Ein brennendes Teelicht (Gegenstandshöhe *G* = 2 cm) befindet sich 10 cm vor der Linse (Abb. 1).
Mit einer weißen Pappe kann sie das Bild 7 cm hinter der Linse auffangen. Es hat eine Bildgröße von 1,4 cm.

Bestimme den Brennpunkt mit einer Strahlenkonstruktion!

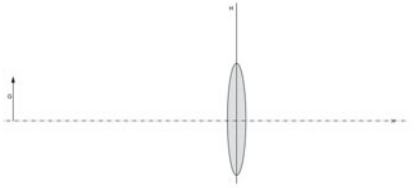


Abbildung 23: Aufgabenblatt 6: Übungen zur Bildkonstruktion [Mat. 9 _ K1] bzw. [Mat. 9_K2] links Aufgaben der TG, rechts Aufgabe der KG

Hierzu operierte die Treatmentgruppe wie zuvor mit der schematischen Darstellung eines leuchtenden Gegenstands, der im Abstand *g = 2f* vor einer Sammellinse steht (vgl. Abbildung 23). Die Schüler wurden nun kognitiv aktiviert, sich vorzustellen, was mit der Größe und Lage des Bildes passiert, wenn man einen Gegenstand auf die Linse zuschiebt bzw. weiter von der Linse wegschiebt. Das Ergebnis dieser Überlegungen sollten die Schüler jeweils in der gegebenen Abbildung durch eine Skizze darstellen (siehe Abbildung 23).

Schließlich wurden die Schüler gebeten, sich den extremen Fall vorzustellen, in dem der Gegenstand sehr weit von der Linse entfernt wurde und Konsequenzen für die Bildweite aufzuzeigen.

Die Kontrollgruppe bearbeitete parallel dazu an zwei Strahlenkonstruktionen. In der ersten Aufgabe konstruierte die Kontrollgruppe das Bild eines vergrößerten selbstleuchtenden Gegenstands; die zweiten Aufgabe enthielt ein Transferproblem: hier waren die Werte der physikalischen Größen des Bildes gegeben. Mit Hilfe der Strahlenkonstruktion sollte der Brennpunkt bestimmt werden (vgl. Abbildung 23).

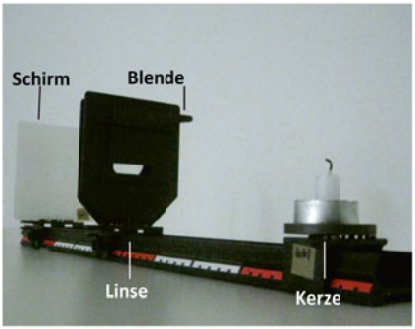
Auch die Kontrollgruppe operierte also mit der schematischen Repräsentation der Strahlenkonstruktion und löste hierzu nicht nur Standardfragestellungen, so dass der Unterschied in den Instruktionen der beiden Bedingungen nicht auf die Bearbeitung von Standardaufgaben versus Problemlöseaufgaben reduziert werden kann. Im Gegensatz zur Treatmentgruppe wurde die Kontrollgruppe jedoch nicht kognitiv aktiviert, die Bildentstehung durch das Operieren mit der Strahlenkonstruktion intern mental zu simulieren.

2.3.3.7 Treatmentvariation in der vierten Unterrichtsstunde

Die Treatmentvariation zielte darauf, die weitverbreitete Schülervorstellung zu überwinden, dass bei der Entstehung reeller Bilder das Bild abgeschnitten werde, wenn man eine Blende vor eine Sammellinse halte. Zur Förderung eines angemessenen Verständnisses dieses physikalischen Settings wurde folgende POE-Sequenz umgesetzt, bei der sich die Schüler mit multiplen Repräsentationen auseinandersetzten (zum Unterrichtsmaterial der Treatmentvariation sei auf Anhang C3 und zur Unterrichtsplanung der Treatmentvariation auf Anhang C4 jeweils auf der Produktseite zu diesem Buch unter www.springer.com hingewiesen).

Die Schüler wurden in der variierten Treatmentbedingung kognitiv aktiviert, sich mit dem konkreten Versuch und unterschiedlichen Repräsentationen des physikalischen Phänomens, das in dem Versuch beobachtet werden kann, auseinander zu setzen (Fotografie des Versuchsaufbaus, verbalen Erklärungen des Phänomens, schematische Repräsentation des Strahlengangs).

Im ersten Schritt (predict) sollten die Schüler vorhersagen, was ein Beobachter auf dem Schirm sieht, wenn man die Kerze anzündet und entsprechend der Abbildungen die obere Hälfte der Sammellinse abdeckt. Der fertige Versuchsaufbau wurde den Schülern vom Lehrer gezeigt. Unterstützend erhielten die Schüler zur Bearbeitung ihrer Vorhersage auf einem Arbeitsblatt folgende Fotografien des Versuchsaufbaus aus unterschiedlichen Perspektiven (vgl. Abbildung 24):



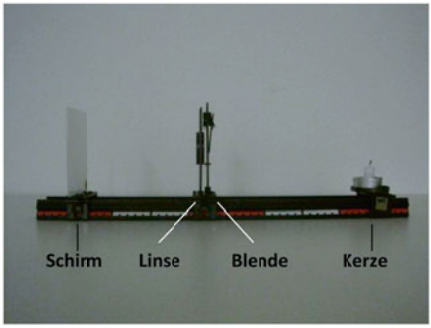


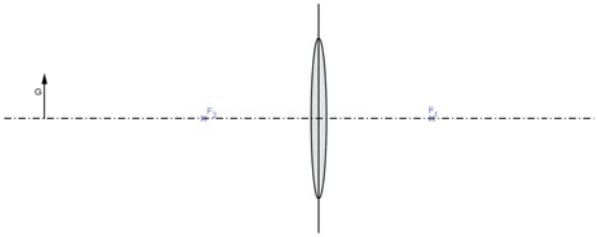
Abbildung 24: Demonstrationsversuch „Abgedeckte Sammellinse“ [Zusatz-1-K1]

Ihre Vermutungen sollten die Schüler bezogen auf die Abbildungen schriftlich auf dem Arbeitsblatt formulieren. Im nächsten Schritt (observe) wurde den Schülern der Versuch demonstriert (vgl. Abbildung 25). Ihre Beobachtungen wurden in der Klasse diskutiert (observe).

Abbildung 25: Ausgang des Demonstrationsversuchs „Abgedeckte Sammellinse“

Im dritten Schritt (explain) bearbeiteten die Schüler ein Arbeitsblatt (vgl. Abbildung 26), in dem die Erklärung des Phänomens in einer schematischen Strahlenkonstruktion dargestellt wurde. Die grafische Lösung ist in Abbildung 27 dargestellt.

Deutung des Versuchsergebnisses



a) Konstruiere das Bild der Kerze (Abb. oben).

b) Zeichne den Lichtstrahl ein, der vom obersten Punkt des Gegenstands ausgeht und gerade noch den oberen Rand der Linse erreicht.

c) Wie ändert sich das Bild, wenn man nur den Bereich des oberen Randes der Linse abdeckt?

_____

_____

d) Warum sieht man ein vollständiges Bild, wenn man die komplette obere Hälfte des Bilde abgedeckt hat (siehe Demonstrationsversuch)? Stelle grafisch dar, was passiert?

e) Grenze durch Einzeichnen ein, welche Lichtstrahlen nach der Abdeckung der Sammellinse noch zur Bildentstehung beitragen.

Abbildung 26: Demonstrationsversuch „Abgedeckte Sammellinse“ [Zusatz-2-K1]

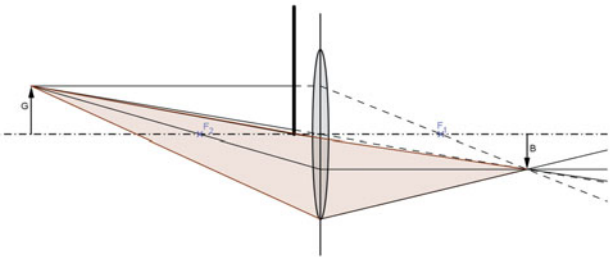


Abbildung 27: Konstruktion des reellen Bildes einer Kerze, welches durch eine Sammellinse entsteht, wobei die obere Hälfte der Linse abgedeckt wurde. Die hervorgehobene Fläche zeigt den Bereich des Lichts, welcher nach Abdeckung der Linse zur Entstehung des reellen Bildes beiträgt.

Die zusätzlich für die Durchführung der POE-Sequenz in der Treatmentgruppe aufgewendete Zeit wurde bei der Erarbeitung der Entstehung virtueller Bilder eingespart. Die Entstehung virtueller Bilder war kein Gegenstand des Leistungstests. Die Kontrollgruppe setzte sich entsprechend wie in Abschnitt 2.3.3.6 Operationalisierung des Treatments in der vierten Unterrichtsstunde dargestellt, ausführlicher mit der Konstruktion des virtuellen Bildes auseinander.

2.3.3.8 Operationalisierung des Treatments in der fünften Unterrichtsstunde

Bevor sich die Schüler mit dem Abbildungsgesetz befassten, bearbeiteten sie ein weiteres Arbeitsblatt zur Vertiefung der Strahlenkonstruktion, in dem auch weitverbreitete Schülervorstellungen thematisiert wurden. Dieses Arbeitsblatt (Arbeitsblatt 7) wurde zum Zweck der Datenerhebung in beiden Gruppen eingesammelt. Zur Ergebnissicherung erhielten die Schüler die Lösung der Aufgaben für ihre Lernunterlagen.

In beiden Gruppen befassten sich die Schüler mit der Strahlenkonstruktion unter den Bedingungen, dass die Linse erstens kleiner als der leuchtende Gegenstand ist und zweitens die obere Hälfte der Linse abgedeckt wird. Die hier adressierten Lernschwierigkeiten betreffen die physikalischen Vorstellungen des Abbildungsvorgangs. Viele Schüler nutzen zur Erklärung der Entstehung des reellen Bildes bei der Sammellinse nicht das Konzept einer „Punkt-zu-Punkt-Abbildung“ (vgl. Wiesner, 1994). Eine der gängigen Vorstellung besteht in diesem Kontext in einer holistischen Erklärung des Abbildungsvorgangs, nach der das Bild als Ganzes durch die Linse zum Schirm geht und dabei in der Linse umgedreht wird (Wiesner, 1994, S. 8). Eine solche holistische Erklärung des Abbildungsvorgangs zeigt sich insbesondere bei den hier verwendeten Abdeckaufgaben.

Unter der Annahme, das Bild werde als Ganzes vom Gegenstand aus durch die Linse auf den Schirm transportiert, ist es nur konsequent anzunehmen, dass ein Teil des Bildes abgeschnitten werde, wenn man eine Blende vor die Linse hält, (Wiesner, 1992b). Hält man eine ringförmige Blende vor die Linse, glauben viele Lernende entsprechend, das Bild werde ringförmig am äußeren Rand abgeschnitten. Wird die Linse zur Hälfte abgedeckt, gehen viele Schüler und auch Studenten davon aus, dass auch das reelle Bild zur Hälfte abgeschnitten werde, einige überlegen sich sogar, welche Hälfte des Bildes (obere versus untere Hälfte) betroffen sei (Goldberg & McDermott, 1987, S. 112; Wiesner, 1994, S. 8). Schüler, die bereits an dieser Stelle über ein adäquates mentales Modell der Lichtausbreitung und der Bildentstehung verfügen, sollten bereits hier erkennen, dass auch

in diesem Fall ein vollständiges Bild entsteht, das in seiner Intensität jedoch abgeschwächt ist. Das Problem in der repräsentationalen Darstellung besteht darin, dass – sowohl im Fall einer kleineren Linse als auch im Fall der abgedeckten Linse – Brennpunktstrahl und Parallelstrahl nicht mehr auf die Linse treffen. Die Lösung besteht darin, zunächst die Hauptebene der Linse zur Strahlenkonstruktion zu verwenden, um herauszufinden, wo das Bild entsteht. Anschließend können diejenigen Strahlen eingezeichnet werden, die auf den oberen und unteren Rand der Linse treffen. Auf diese Weise kann der Bereich angegeben werden, in dem das Licht, welches von dem leuchtenden Gegenstand ausgeht, zur Entstehung des reellen Bildes beiträgt.

Die Treatmentgruppe bearbeitete beide Aufgaben (Bildkonstruktion, wenn Linse kleiner als der leuchtender Gegenstand; Bildkonstruktion, wenn die oberen Linsenhälfte abgedeckt ist) in zwei Schritten (siehe Abbildungen 28 und 29).

- Im ersten Schritt bestand die Aufgabe darin, das Bild unter den gegebenen Bedingungen zu konstruieren. Diese Aufgabenstellung bezweckte, die Schüler ihre Vorstellung des Abbildungsvorgangs in einer externen Repräsentation reflektieren zu lassen. Da die Arbeitsblätter der Schüler eingesammelt wurden, ergibt sich die Möglichkeit, den Lernstand der Schüler auszuwerten, um zu sehen, ob sich die Instruktionen als fruchtbar erweisen, welche diesen Lernschwierigkeiten vorbeugen sollen. So war etwa in Aufgabenblatt 3 bereits durch das Arbeiten mit der Strahlenkonstruktion thematisiert worden, dass alles Licht zur Bildentstehung beiträgt, das auf die Sammellinse fällt.
- In einem zweiten Schritt erhielten die Schüler eine graphische Teillösung. Sie wurden nun aufgefordert, die Teillösung zu ergänzen und anhand der Abbildung zu erläutern, warum es möglich ist, auch dann noch ein Bild zu konstruieren, wenn der Gegenstand kleiner als die Linse ist bzw. wenn die obere Hälfte der Linse abgedeckt worden war (siehe Abbildung 28 und Abbildung 29).

Die Schüler sollten in mehreren Hinsichten durch diese Aufgabenstellung kognitiv aktiviert werden:

- Durch den Abgleich der eigenen Vorstellung mit der Teillösung sollten die Schüler ihre Sichtweise reflektieren. Die Aufgabe zielte darauf, ihre Lösung des physikalischen Problems zu prüfen und die eigenen Verarbeitungsschritte entsprechend anzupassen.
- Die Vervollständigung der Strahlenkonstruktion war als repräsentationales Fading-out gedacht, bei dem die Schüler erneut die Bildkonstruktion durch-

denken sollten, sofern sie mit der Lösung der Aufgabe Schwierigkeiten hatten.

- Die schriftliche Erklärung, warum es möglich ist, ein Bild unter den gegebenen Bedingungen (Linse < Gegenstand) zu konstruieren, erforderte das Erstellen von zwei Repräsentationen in unterschiedlichen Formaten (Bild und Text), die aufeinander bezogen werden mussten. Bei der Abdeckaufgabe wurde zudem gefragt, wie sich das Abdecken auf das Bild auswirkt. Hier galt es aus der Konstruktion und der Beschreibung eine Schlussfolgerung über Auswirkungen auf das reelle Bild (Ebene des Phänomens) zu beziehen.

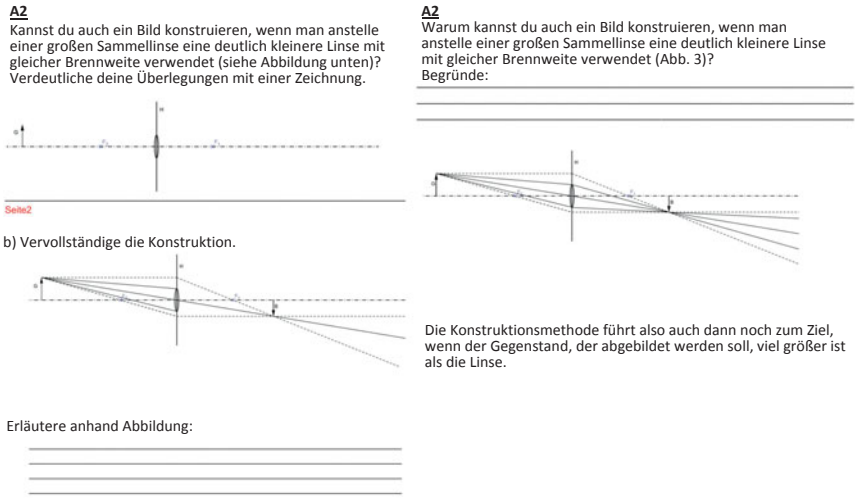
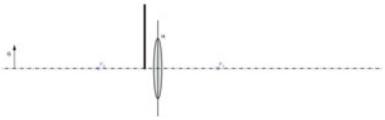


Abbildung 28: Aufgabenblatt 7 Wiederholung zur Bildkonstruktion[Mat. 10_K1], A2 bzw. [Mat 10_K2], A2 links Aufgabe der TG, rechts Aufgabe der KG

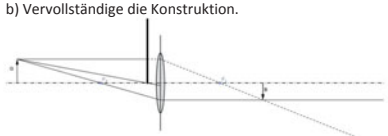
In der Kontrollgruppe wurde die gleiche Schülervorstellung adressiert.

Im Gegensatz zur Treatmentgruppe erhielt die Kontrollgruppe die Abbildung inklusive vollständiger Strahlenkonstruktion und sollte die Frage beantworten, warum es jeweils möglich ist, unter den gegebenen Bedingungen (Linse < Gegenstand; Abdeckung der oberen Linsenhälfte), das Bild zu konstruieren (vgl. Abbildungen 29). Im Gegensatz zur Treatmentgruppe wurde die Kontrollgruppe nicht aufgefordert, mit der Konstruktion selbst zu operieren. Die Aufgabe, beide Repräsentationen aufeinander zu beziehen, bestand zwar auch in der Kontrollgruppe; die zentrale graphische Repräsentation war der Kontrollgruppe jedoch vorgegeben. Somit wurde die Kontrollgruppe nicht aktiviert, ihre eigenen Vorstellungen extern darzustellen, diese auf ihre Zweckmäßigkeit hin zu prüfen oder die eigenen Verarbeitungsschritte entsprechend anzupassen.

A3
a) Kannst Du noch das Bild einer Kerze (hier als Pfeil dargestellt) konstruieren, wenn man entsprechend der Abbildung die obere Hälfte der Sammellinse abdeckt? Verdeutliche Deine Überlegungen mit einer Zeichnung (Abb. 5).



b) Vervollständige die Konstruktion.



Was passiert mit dem Bild, wenn man wenn man entsprechend der Abbildung die obere Hälfte der Sammellinse abdeckt? Erläutere anhand der Abbildung oben:

A3
Was passiert mit dem Bild auf dem Schirm, wenn man entsprechend der Abbildung die obere Hälfte der Sammellinse abdeckt? (Abb. 4)?

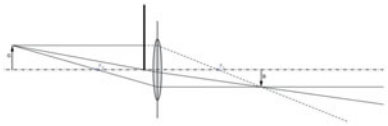


Abbildung 29: Aufgabenblatt 7: Wiederholung zur Bildkonstruktion [Mat_10K1], A3 bzw. [Mat. 10_K2] A3, links Aufgabe TG, rechts Aufgabe der KG

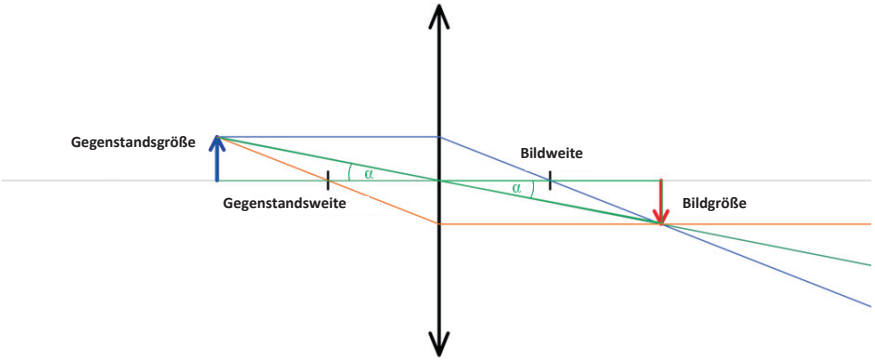


Abbildung 30: Aufgabenblatt 8: Die Abbildungsgleichung bei der Sammellinse [Mat.12_K1] bzw. [Mat. 12_K2] geometrische Veranschaulichung zur Herleitung der Abbildungsgleichung

Im zweiten Teil der Stunde wurde die Abbildungsgleichung eingeführt und verbal sowie geometrisch veranschaulicht (vgl. Abbildung 30). Das hier eingesetzte Arbeitsblatt war zuvor in der Pilotstudie von Scheid (2013) verwendet worden. Schülervorstellungen wurden in diesem Kontext nicht thematisiert.

Betrachtet man die beiden Dreiecke, welche die Gegenstandsgröße bzw. Bildgröße, die Gegenstandsweite bzw. die Bildweite und den Mittelpunktstrahl

als Seiten haben, können hieraus unter Anwendung des geometrischen Satzes zur Ähnlichkeit von Dreiecken eine allgemeine Formel über die Seitenverhältnisse der Dreiecke abgeleitet werden. Da die Ähnlichkeit von Dreiecken jedoch gemäß Lehrplan des Landes Rheinland-Pfalz im Fach Mathematik nicht zuvor im Mathematikunterricht behandelt worden war (vgl. Ministerium für Bildung, Wissenschaft, Jugend und Kultur. Rheinland-Pfalz. Rahmenlehrplan Mathematik, 2007), wurden die Schüler in beiden Gruppen schrittweise zu diesem Zusammenhang hingeführt.

Die Treatmentgruppe erhielt hierzu mehrere aufeinander aufbauende Aufgaben:

- In einem ersten Schritt sollten die Schüler, wie zuvor geübt, den Strahlengang in der gegeben Abbildung durch das Einzeichnen des Mittelpunktstrahls ergänzen. In der folgenden Abbildung waren alle irrelevanten Informationen ausgeblendet, so dass nur die für die Herleitung der Gleichung relevanten Dreiecke zu sehen sind. Diese sollten die Schüler mit den entsprechenden physikalischen Größen korrekt beschriften und sich Gedanken über die Winkelverhältnisse der Dreiecke machen.
- Im nächsten Schritt wurden die Schüler aufgefordert sich zu veranschaulichen, dass beide Dreiecke in ihren Seitenverhältnissen übereinstimmen; ihre Schlussfolgerung über die Seitenverhältnisse sollten sie in einem kurzen Satz schriftlich festhalten.
- In den letzten Schritten wurde von den Schülern gefordert, diesen Zusammenhang über den Zwischenschritt einer Wortgleichung als Formel auszudrücken.

Der Kontrollgruppe wurde parallel dazu der gleiche Zusammenhang unter Verwendung der gleichen Repräsentationen in einem Lehrervortrag erläutert. Sie erhielt somit die gleichen Informationen, jedoch ohne gezielte kognitive Aktivierung zu den jeweiligen Bearbeitungsschritten.

Folgende Aspekte der Aufgabenstellung zielten auf eine kognitive Aktivierung beim Operieren mit den Repräsentationen in der Treatmentgruppe im Vergleich zur Kontrollgruppe:

- Die Treatmentgruppe war gefordert, die relevanten Informationen selbst in den gegebenen graphischen Repräsentationen zu ergänzen.
- Die Aufgabenstellung, sich die Übereinstimmung der Seitenverhältnissen mental zu veranschaulichen und die Schlussfolgerung zu verbalisieren, zielte auf die Bildung eines mentalen Modells des mathematischen Zusammenhangs;

aus diesem mentalen Modell können die relevanten Schlussfolgerung im Idealfall abgelesen werden.
- Durch das schriftliche Festhalten der Schlussfolgerung und die Übersetzung der Schlussfolgerung in die Sprache abstrakter Formeln sollten die Schüler lernen, die Schlussfolgerungen aus der grafischen Repräsentation sowohl in eine sprachliche als auch in eine mathematische Repräsentation zu übersetzen.

Im Anschluss wandten beide Gruppen die Abbildungsgleichung in einer Übungsaufgabe an. Die Aufgabenformulierung und der Rechenaufwand waren in beiden Gruppen parallel gehalten. Die Treatmentgruppe wurde jedoch aufgefordert, das Ergebnis ihrer Rechnung anhand einer gegeben Strahlenkonstruktion, die im Originalmaßstab ausgeteilt wurde, zu beziehen, um die Verbindung zwischen Abbildungsgleichung und Strahlenkonstruktion zu festigen.

2.3.3.9 Operationalisierung des Treatments in der sechsten Unterrichtsstunde

Die letzte Unterrichtsstunde steuerte darauf hin, die gelernten Inhalte durch Übungen zur Bildkonstruktion und zum Abbildungsgesetz zu vertiefen (Aufgabenblatt 9). Hierzu bearbeiteten die Schüler zwei Aufgaben mit Teilfragestellungen. In der ersten Aufgabe wurde der Treatment- und auch der Kontrollgruppe die graphische Repräsentation einer Kerze (dargestellt durch einen Pfeil) vor einer Sammellinse präsentiert. Beide Gruppen sollten sich überlegen, ob das Bild der Kerze auch bei einer Versuchsanordnung ohne Schirm gesehen werden kann. Die hier adressierte Schülervorstellung wurde in der (bereits in vorigen Kapiteln erläuterten) Studie von Goldberg und McDermott (1987) beschrieben.

Zur Erinnerung: Goldberg und McDermott (1987) zeigten Studenten einen Versuchsaufbau, bei dem eine Glühbirne, eine Linse und ein Schirm hintereinander auf einer optischen Bank montiert waren. Die Studenten wurden gefragt, wo das Bild wäre, wenn man den Schirm entfernt und sie frei um den Versuchsaufbau im Raum herumgehen könnten. Nur wenige Studenten waren in der Lage zu erkennen, dass sich das Bild an der gleichen Position befindet wie der Schirm. Die übrigen Studenten gaben unter anderem die Erklärungen ab, das Bild sei auf oder in der Linse. Verbreitet war auch die Vorstellung, dass ein Bild nur mit Hilfe eines Schirms gesehen werden kann, wobei die Linse das Bild quasi einrahme (vgl. ebd., S. 114).

Die Herausforderung bei dieser Aufgabe besteht darin, zu erkennen, dass sich die Lichtstrahlen, die ein reelles Bild formen, im freien Raum treffen. Unter

der Voraussetzung, dass die Schüler ein kohärentes mentales Modell des Abbildungsvorgangs gebildet haben, sollten sie in der Lage sein, auch in dieser Aufgabe die korrekte Schlussfolgerung zu ziehen. Die Aufgabenstellung in Treatment- und Kontrollgruppe unterschied sich nun wie folgt (vgl. Abbildung 31): In der schematischen Repräsentation des Strahlengangs, welcher der Treatmentgruppe vorgelegt wurde, waren zusätzlich drei Beobachter an unterschiedlichen Positionen platziert.

Die Treatmentgruppe wurde gefragt, an welcher Stelle der Beobachter A, B, oder C ein scharfes Bild einer Kerze sehen kann, wenn man …

1. einen intransparenten Schirm an der Position S aufstellt,
2. den intransparenten Schirm gegen einen transparenten Schirm austauscht oder
3. den Schirm entfernt.

Die Schüler wurden gebeten, ihre Antworten zu begründen. Durch die Fragestellung wurde die Treatmentgruppe kognitiv aktiviert, sich mental in die Position der verschiedenen Beobachter hineinzuversetzen. Die Überlegungen, die zur korrekten Lösung führen, können zudem grafisch durch eine entsprechende Skizze veranschaulicht werden.

Im Fall des intransparenten Schirms können nur die Beobachter A und B das Bild sehen, bei Verwendung eines intransparenten Schirms sehen alle Beobachter das Bild, da das Licht auf dem Schirm in alle Richtungen streut und im letzten Fall (kein Schirm) kann nur Beobachter A das Bild sehen, da nur er im Strahlengang steht. Mit den ersten beiden Fällen (intransparenter und transparenter Schirm) wurde auch implizit nochmals das Konzept der Streuung und die physikalische Sehtheorie (vgl. Wiesner, 1986) thematisiert, weil die Schüler erkennen mussten, dass beleuchtete Gegenstände (hier der Schirm) Licht abstrahlen und dieses Licht in die Augen des Beobachters fallen muss, wenn der Beobachter das Bild sehen kann.

Die Kontrollgruppe wurde im Gegensatz zur Treatmentgruppe direkt gefragt, ob es möglich sei, das Bild der Kerze bei einer Versuchsanordnung ohne Schirm zu sehen und ihre Antwort zu begründen (vgl. Abbildung 8). Die Kontrollgruppe erhielt über die Adressierung der Schülervorstellung hinaus, die an sich kognitiv anspruchsvoll ist, also keine zusätzliche kognitive Aktivierung, mit der gegebenen schematischen Repräsentation der Strahlenkonstruktion mental und zeichnerisch zu operieren.

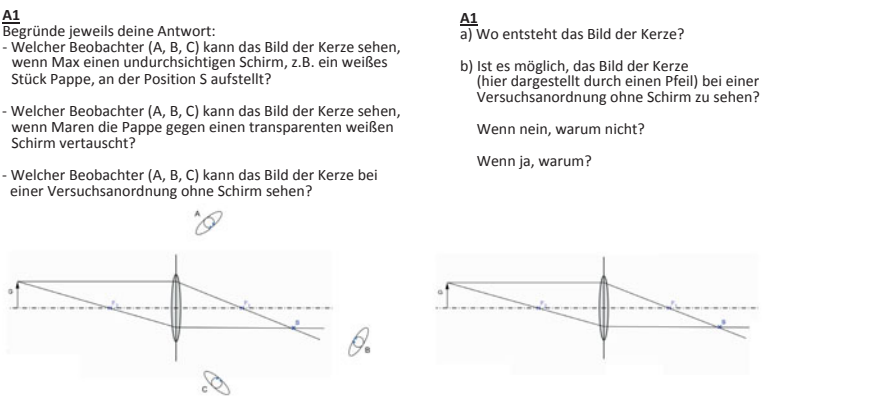


Abbildung 31: Aufgabenblatt 9 Übungen zur Abbildungsgleichung und zur Bildkonstruktion [Mat. 13_K1], A1 bzw. [Mat. 13_K2], A1 links Aufgabe der TG, rechts Aufgabe der KG

Die letzte Übungsaufgabe der Unterrichtseinheit war auf eine Vertiefung der Strahlenkonstruktion angelegt.

Beide Gruppen erhielten zunächst die gleiche graphische Repräsentation, bei der eine Kerze (dargestellt durch einen Pfeil) vor einer Sammellinse steht. Beide Gruppen sollten eine Strahlenkonstruktion durchführen, die von der üblichen Vorgehensweise abweicht.

In der Treatmentgruppe waren hierzu zwei Punkte P und Q auf der Hauptebene (Linsenmitte) angegeben. Die Aufgabe bestand darin, jeweils einen Strahl zu konstruieren, der von der Kerzenspitze ausgeht und in dem Punkt P bzw. in dem Q auf der Linsenmitte auftrifft (vgl. Abbildung 32). Die Lösung der Aufgabe umfasst zwei Bearbeitungsschritte:

- Im ersten Schritt wird wie üblich das reelle Bild unter Verwendung der ausgezeichneten Strahlen eingezeichnet.
- In einem zweiten Schritt kann dann ausgehend von der Konstruktion des Bildes der Verlauf der Strahlen angegeben werden, die am Punkt P bzw. Q auf der Linsenmitte auftreffen und ebenfalls zur Entstehung des Bildes beitragen.

Diese Aufgabe zielte darauf, die Schüler kognitiv zu aktivieren, die Strahlenkonstruktion erneut zu durchdenken und sich zu überlegen, dass außer den ausgezeichneten Strahlen noch mehr Licht zur Bildentstehung beiträgt, was sich wiederum zeichnerisch darstellen lässt.

Die Kontrollgruppe konstruierte parallel hierzu zwei Gegenstandspunkte P und Q, die in der gegebenen Zeichnung eingetragen waren (vgl. Abbildung 32). Sie wurde jedoch nicht kognitiv aktiviert, die Strahlenkonstruktion unter der Perspektive zu durchdenken, dass außer den ausgezeichneten Strahlen auch andere Lichtbündel im zeichnerischen Modell dargestellt werden können.

A2
Konstruiere das Bild der Kerze (hier dargestellt durch einen Pfeil) und bestimme die Bildgröße durch Konstruktion (Abb. 2)!
Hinweis: Die Linse ist durch die Linsenmitte (H) dargestellt.

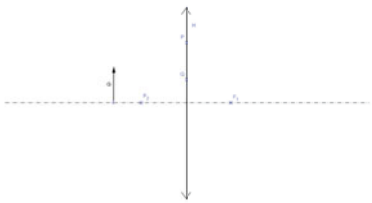


Abb. 2

b) Bestimme den Abbildungsmaßstab

c) Welche Abbildungseigenschaften treffen zu:

o seitenrichtig o seitenverkehrt
o vergrößert o verkleinert

d) Zeichne folgende Strahlen ein:

- Strahl, der von der Kerzenstpitze ausgeht und in dem Punkt P auf der Linsenmitte auftrifft
  und
- Strahl, der er von der Kerzenstpitze ausgeht und in dem Punkt Q auf der Linsenmitte auftrifft

A2
a) Konstruiere das Bild des Pfeils und bestimme die Bildgröße durch Konstruktion (Abb. 2)!
Hinweis: Die Linse ist durch die Linsenmitte (H) dargestellt.

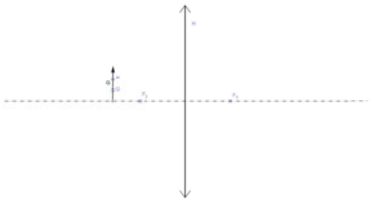


Abb. 2

b) Bestimme den Abbildungsmaßstab

c) Welche Abbildungseigenschaften treffen zu:

o seitenrichtig o seitenverkehrt
o vergrößert o verkleinert

d) Konstruiere die Bildpunkte zu den Punkten P und Q des Gegenstandes

Abbildung 32: Aufgabenblatt 9 Übungen zur Abbildungsgleichung und zur Bildkonstruktion [Mat. 13_K1] bzw. [Mat. 12 _K2], A2 links Aufgabe der TG, rechts Aufgabe KG

2.3.4 Variablen und Erhebungsinstrumente

2.3.4.1 Überblick

Tabelle 8 zeigt einen Überblick über die in der Hauptstudie erhobenen abhängigen Variablen und Kovariaten und ihre Operationalisierung.

Tabelle 8 Überblick über Variablen und die Erhebungsinstrumente

Variable	Erhebungsinstrument
abhängige Variablen	
Wissen und Problemlösen beim Umgang mit Repräsentationen in der Strahlenoptik	Leistungstest (25 min.): Im Rahmen der Studie entwickeltes Testinstrument, drei Messzeitpunkte
Konzeptuelles Grundverständnis in Strahlenoptik	Konzepttest (15 min.): Im Rahmen der Studie entwickeltes Testinstrument, drei Messzeitpunkte
Motivation der Schüler	Fragebogen (5 min.), Basierend auf Hoffmann, Häußler & Peters-Haft, 1997; entnommen aus Kuhn (2008, S. 305-308)
Kovariaten	
Relevante kognitive Fähigkeiten Verbale Fähigkeiten Räumliches Denken (figural) Logisches Schlussfolgern (figural)	Intelligenz-Struktur-Test (I-S-T) 2000 R: drei Subskalen (45 min.), Satzergänzung, Würfelaufgaben und Matrizen, einmalige Erfassung
Vorleistung in Mathematik, Physik, Deutsch	Jeweilige letzte Zeugnisnote
Weitere mögliche Einflussfaktoren	
Geschlechtszugehörigkeit	Dummy – Kodierung (0 = weiblich, 1 = männlich)
Schultyp	Dummy – Kodierung (0 = IGS, 1 = Gymnasium)
Klassengröße	Anzahl teilnehmender Schüler je Schulklasse
Einflussfaktoren bei der Anwendung des Lehrmaterials	
Ablauf der Unterrichtsstunden	Lehrernotizheft (Items entnommen aus Helmke et al., 2010), einmalige Erhebung nach jeder Unterrichtsstunde
Engagement des Lehrers aus Schülersicht	Teilfragen des obigen Motivationsfragebogens, Skala aus: Seidel, Prenzel, Duit & Lehrke (2003, S. 366), drei Messzeitpunkte

2.3.4.2 Leistungstest

Zur Erhebung der abhängigen Variablen Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen wurde der Leistungstest der Pilotstudie

von Jochen Scheid in Zusammenarbeit mit der Autorin für die Studie grundlegend überarbeitet: acht Aufgaben kamen neu hinzu.

Das in dieser Studie verwendete Testinstrument bestand aus 14 Aufgaben und erfasste Wissen und Problemlösen im Bereich der Strahlenoptik durch Aufgaben, die den Umgang mit fachspezifischen Repräsentationen erfordern (siehe Tabelle 9 zum Überblick über die repräsentationalen Anforderungen). Die Testdauer betrug sowohl im Prä- als auch im Posttest 25 Minuten. Maximal konnten jeweils 42 Punkte je Test erreicht werden. Prä-, Post- und Follow-up Test waren identisch. Der Test umfasste folgende Aufgabentypen:

- Repräsentationsbezogene Aufgaben, die nur korrekt gelöst werden konnten, wenn das Denken der Schüler nicht durch spezifische Schülervorstellungen geleitet wurde (Aufgaben 1, 4a und 4b).[23] Aufgaben dieses Typs zielten darauf, Wissen und Problemlösen beim Umgang mit Repräsentationen in unterschiedlichen Formaten im Hinblick auf die Überwindung spezifischer Schülervorstellungen zu erfassen. Sie waren mit den Aufgaben 1, 2a und 2b des Pilotstudienvortests identisch. Beispiel: „Wie ändert sich das Bild des Gegenstandes, wenn man eine Lochblende vor die Linse stellt? Begründe Deine Antwort durch eine Zeichnung. Verwende hierfür die gegebene Abbildung (unten)."
- Aufgaben, bei denen die Schüler mit der schematischen Repräsentationsform des Strahlengangs operierten (Aufgabe 2 Teilbereich „Zeichnung", Aufgabe 4a und 4b - Teilaufgabe Zeichnung)[23]. Beispiel: „Zeichne den Strahlengang des folgenden Versuchsaufbaus".
- Aufgaben, in denen die Schüler Repräsentationen in depiktional schematischer Form in eine deskriptiv verbale Repräsentationsform übersetzen (2a, b; 4b - Teilaufgabe-Erklärung, 6)[23] oder mit einer deskriptiven Repräsentation operieren sollten (3, 7). Beispiel: „Ein Gegenstand befindet sich vor einer Linse. Die Gegenstandsweite liegt zwischen einfacher und doppelter Brennweite der Linse. Wie groß ist das Bild verglichen mit der Gegenstandsgröße? Wie groß ist die Bildweite verglichen mit der Brennweite?"
- Aufgaben, in denen die Schüler Repräsentationen auf Basis einer depiktional realistischen Repräsentation Schlussfolgerungen ziehen und diese verbal formulieren sollten. Beispiel: „Ein gewöhnliches Zimmerfenster wird durch eine Sammellinse auf einer Wand in einem Raum abgebildet. Um welchen Bildfall handelt es sich?"

23 Mehrfachzuordnung der Aufgaben 4a und 4b: Teilaufgaben Zeichnung und Erklärung

Die beiden letztgenannten Aufgabentypen waren daraufhin angelegt, das fachliche Wissen im Umgang mit Repräsentationen und die repräsentationale Kohärenz zu erfassen.

Tabelle 9 Übersicht über die repräsentationalen Anforderungen der Aufgabentypen im Leistungstest

Aufgabe	Thematisierte Schülervorstellung	Repräsentationale Anforderung	Punkt
A 1	Abdeckaufgabe	Ein mentales Modell der Situation auf Basis von Hinweisen in depiktionaler Form entwickeln, aus diesem Modell Schlüsse ziehen und diese verbal begründen (identisch mit Aufgabe 1 der Pilotstudie).	4
A 2 Zeichnung	-	Entwicklung einer sachlich korrekten und vollständigen schematischen Repräsentation (Konstruktion des Strahlengangs mit vollständiger Beschriftung).	7
Teil a)	-	Auf Basis der angefertigten Konstruktion logische Schlüsse ziehen und diese verbal ausdrücken (Operation mit der selbst entwickelten Repräsentation: mentale Simulation der Bildentstehung).	4
Teil b)	-	Auf Basis der angefertigten Konstruktion logische Schlüsse ziehen und diese verbal ausdrücken.	4
A 3	-	Ein mentales Bild der Situation aufbauen oder eine Skizze anfertigen und dann auf Basis dieser bildlichen internalen oder externalen Repräsentation Schlussfolgerungen ziehen und diese verbal bzw. unter Verwendung der Abkürzung der entsprechenden physikalischen Formel ausdrücken.	2
A 4 Teil a)	Abdeckaufgabe	Eine schematische Repräsentation entwickeln und mit dieser operieren und physikalische Schlussfolgerungen (Konsequenzen auf der Ebene der Phänomene) ziehen (identisch mit Aufgabe 2a der Pilotstudie).	5
Teil b) mit Erklärung	Abdeckaufgabe	Eine schematische Repräsentation entwickeln, mit dieser operieren und physikalische Schlussfolgerungen (Konsequenzen auf der Ebene der Phänomene) ziehen und diese begründen (identisch mit Aufgabe 2b der Pilotstudie).	3 + 1,5
A 5 Teil a)	-	Auf Basis einer realistisch depiktionalen Abbildung ein Phänomen kategorisieren.	1
Teil b)	-	Auf Basis einer realistisch depiktionalen Abbildung ein Phänomen kategorisieren.	1
Teil c)-1	-	Die Entscheidungen der Kategorisierung 5a) und 5b) verbal	1
Teil c)-2		begründen.	1
A 6	-	Eine gegebene Repräsentation korrigieren und an die Anforderungen der Fragestellung anpassen. Den Zweck einer bildlich-schematischen Repräsentationen erkennen (metakognitiver Aspekt). Aus einer schematisch bildlichen Repräsentation mathematisch (logische) Schlussfolgerungen ziehen. Bildlich-schematische Repräsentationen in verbale Repräsentationen übersetzen, eigene Schlussfolgerungen begründen.	3
A 7	-	Mit einer mathematischen Repräsentation operieren (identisch mit Aufgabe 7 der Pilotstudie).	4,5
			∑ 42

Ebenso wie der Test der Pilotstudie orientierte sich auch das Erhebungsinstrument der Hauptstudie am Lehrplan für Gymnasien in Rheinland-Pfalz der Jahrgangsstufe 8. Eine ausführlichere Beschreibung des Tests kann in der Dissertationsschrift von Scheid (2013, S. 139 ff.) nachgelesen werden. Der vollständige in dieser Arbeit verwendete Test ist in Anhang C5 auf der Produktseite zu diesem Buch unter www.springer.com enthalten.

2.3.4.3 Konzepttest

Der Konzepttest wurde für die Hauptstudie wie folgt überarbeitet, eine tabellarische Übersicht zur Überarbeitung findet sich in Tabelle 7 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com:

- Items, die Konzepte zur Bildentstehung beim Hohlspiegel erfragten, wurden weggelassen, da die Bildentstehung an der Sammellinse Thema der Untersuchung war.
- Hinzu kamen drei Items zum Thema Streuung und physikalische Sehvorstellung, weil das Verständnis der Strahlenkonstruktion auf diesen Konzepten basiert sowie zwei Items zum virtuellen Bild, dessen physikalische Erklärung im Gegensatz zur Pilotstudie in der Hauptstudie mit aufgenommen wurden.[24]
- Der Test wurde homogener gestaltet. Jedes Item enthielt drei Distraktoren und eine korrekte Antwort, mit Ausnahme von Item 5 und 8, in denen aus sachlichen Gründen zwei korrekte Antworten bestanden.
- Items mit negativer Trennschärfe im Posttest der Pilotstudie wurden entfernt oder überarbeitet.

Die Punktevergabe der Items erfolgte analog zur Pilotstudie. Entsprechend konnten je Item maximal 0, 1 bis 2 Punkte erzielt werden. Die maximal erreichbare Punktzahl lag bei 21 x 2 Punkten = 42 Punkte. Der vollständige Test befindet sich in Anhang C6 auf der Produktseite zu diesem Buch unter www.springer.com. Folgende Grundkonzepte wurden in dem Test erfasst (vgl. Tabelle 10)

24 Aus zeitlichen Gründen wurde in der Pilotstudie auf die Erklärung virtueller Bilder am Hohlspiegel verzichtet. Da in der Hauptstudie die Bildentstehung umfassend gemäß des Lehrplans eingeführt werden konnte und die Thematisierung virtueller Bilder auch in Hinblick auf die Förderung des konzeptuellen Verständnisses reizvoll ist, wurde dieses Thema (wenn auch nicht schwerpunktmäßig) mit aufgenommen.

Tabelle 10 Erfasste Grundkonzepte und zugeordnete Items des Konzepttests

korrekte Lösung: wiss. Konzept	**Distraktoren**: (wiss. inadäquate) Schülervorstellungen	**Item**
Lichtausbreitung	Licht wird mit seiner Quelle / seinen Wirkungen / mit einem Zustand gleichgesetzt.	1, 9, 10
Sekundäre Lichtquellen und Physikalische Sehvorstellung: Streuung	Beleuchtete Gegenstände wie Tische, Bücher oder Bilder strahlen kein Licht ab / beleuchtete Gegenstände werfen kein Licht zurück.	3, 9, 10
Physikalische Sehvorstellung	Bei Lichtquellen mit geringer Intensität gelangt kein Licht mehr ins Auge.	2
Wissenschaftstheoretisch adäquates Modellverständnis	Licht wird als Flüssigkeitsstrom beschrieben, der in Bewegung ist, sich aber auch in Ruhelage befinden kann. Verwechslung von Lichtbündel und Lichtstrahl.	4, 4
Entstehung reeller Bilder bei der Sammellinse	Eine Sammellinse macht das Licht größer Das reelle Bild bei der Sammellinse entsteht durch Spiegelung oder Reflexion. Holistische Konzeption des Abbildungsvorgangs: das Bild geht als Ganzes durch die Linse zum Schirm und wird dabei in der Linse umgedreht. Bei Abdeckung der oberen bzw. unteren Hälfte der Linse wird entsprechend die obere bzw. untere Hälfte des Bildes angeschnitten. Bei Verwendung einer ringförmigen Blende werden die Bildränder kreisförmig abgeschnitten.	5, 11, 15a), 15b), 15c)
Entstehung reeller Bilder bei der Sammellinse:	Fehlerhafte bildliche Repräsentation der Strahlenkonstruktion. Fehlerhaftes Verständnis der Funktionsweise von Sammellinsen. Schwierigkeiten beim Verständnis des Bildortes und der Funktion des Schirms: Vorstellung, das Bild sei auf der Linse; nur wenige Lernende erkennen, dass sich das Bild ohne Schirm an der gleichen Position befindet wie der Schirm.	13, 8, 15d), 15e), 15f)
Ort und Lage des Spiegelbilds	Das Spiegelbild liegt auf der Spiegeloberfläche.	6,
Gerichtete Reflexion (Planspiegel)	Der Spiegel werde als ein Gegenstand aufgefasst, der das Spiegelbild zum Betrachter zurückwirft. (Licht muss also nicht aus Richtung des Spiegels ins Auge fallen, damit das Spiegelbild wahrgenommen werden kann.)	7
Virtuelle Bilder bei der Sammellinse	Probleme beim Verständnis der Entstehung (und Lage) virtueller Bilder.	12, 16
	Punkte je Item: 2, Maximalpunktzahl 42	$\sum_{Items}^{21}$

2.3.4.4 Motivationsfragebogen

Der Motivationsfragebogen bestand aus Items validierter Erhebungsinstrumente und umfasste die Skalen: Selbstkonzept, intrinsische Motivation, (wahrgenommenes) Lehrer-Engagement aus Schülersicht und die Skala „Folgen und Folgenanreize – Gute Noten“. Die Skala „Lehrer-Engagement aus Schülersicht“ (LES) wurde eingesetzt, um möglich Unterschiede zwischen Treatment- und Kontrollgruppe zu erfassen, welche die Lernleistung beeinflussen, bei der Entwicklung des Treatments jedoch nicht intendiert waren. Für den ersten Messzeitpunkt zielt die Skala auf die Erfassung des Engagements ihres Lehrers aus Sicht der Schüler. Hier sollten zwischen Treatment und Kontrollgruppe möglichst keine Unterschiede bestehen. Zum zweiten Messzeitpunkt interagiert die Bewertung mit den eingesetzten Materialien, welche sich in Treatment- und Kontrollgruppe unterscheiden. Wünschenswert wäre, dass Schüler der Treatmentgruppe in der Postmessung zu einer ähnlichen Einschätzung kommen, wie Schüler der Kontrollgruppe.

Insgesamt bestand der Fragebogen aus 28 Items, die auf einer sechsstufigen Skala mit den Polen „trifft gar nicht zu“ (1) bis „trifft voll und ganz zu“ (6) bewertet werden konnten.

Der Posttest wurde um 6 selbstformulierte Items ergänzt, welche das Selbstkonzept in Bezug auf die erworbenen Lerninhalte und den Umgang mit zentralen Repräsentationen erfassten. Diese Items wurden mit dem Kürzel FI (Forschungsitem zur künftigen Auswertung) versehen und im Rahmen dieser Arbeit nicht ausgewertet. Die verwendeten Fragebögen sind in Anhang C7 auf der Produktseite zu diesem Buch unter www.springer.com zu finden. Der Fragebogen setzte sich wie folgt zusammen (vgl. Tabelle 11).

Tabelle 11 Skalen des Motivationsfragebogens

Skala	Items-Anzahl	Beispiel	Quelle der Items
Selbstkonzept (SK)	10	„Der Unterrichtsstoff in Physik ist für mich verständlich.“	Basierend auf Hoffmann, Häußler & Peters-Haft, 1997; entnommen aus Kuhn (2008, S. 305-308)
Intrinsische Motivation bzw. Engagement (IE)	8	„In meiner Freizeit beschäftige ich mich auch über die Hausaufgaben hinaus mit Themen, die mit Physik zu tun haben.“	Basierend auf Hoffmann, Häußler & Peters-Haft, 1997; entnommen aus Kuhn (2008, S. 305-308)
Lehrer-Engagement aus Schülersicht (LES)	4	„Mein Physiklehrer wirkt begeistert im Physikunterricht.“	Aus Seidel et al. (2003, S. 366)
Folgen und Folgenanreize – Gute Noten (GN)	6	„In Physik viel zu können und gut zu sein ist für mich wichtig, weil ich gute Noten bekommen möchte.“	Rheinberg & Wendland (2003, 2004)

2.3.4.5 Erfassung der Kovariaten

Zur Erfassung relevanter kognitive Fähigkeiten wurden Teilskalen des I-S-T 2000 R eingesetzt (Liepmann, Beauducel, Brocke & Amthauer, 2007). Ziel war es, schlussfolgerndes Denken im verbalen und figuralen Bereich zu erfassen (siehe auch Tabelle 12 und Abbildung 33). Aus Zeit- und Kostengründen musste ein Instrument gewählt werden, das eine Gruppentestung ermöglicht. Für ein tiefergehendes Verständnis der Bildentstehung bei der Sammellinse sollten gerade diese Fähigkeiten eine zentrale Rolle spielen. Wie zuvor ausgeführt, besteht eines der Lernziele in der Förderung des kompetenten Umgangs mit multiplen Repräsentationen, wobei vor allem schematische Repräsentationen und verbale Repräsentationen wechselseitig verknüpft werden. Die mathematische Ebene spielt beim Verständnis der Abbildungsgleichung eine Rolle. Der Schwerpunkt wurde in den entwickelten Lernmaterialien auf die geometrische Herleitung gelegt. Zur Bildung eines adäquaten mentalen Modells der Bildentstehung ist zudem die Verknüpfung mit der realen Experimentiersituation und damit das Verständnis der physikalischen Zusammenhänge im dreidimensionalen Raum wichtig. Daher fiel die Wahl auf Skalen zur Erfassung des figural-räumlichen und figural-logischen Schlussfolgerns. Um die Schüler und auch die teilnehmende Schulen

nicht übermäßig zu belasten, wurde die Zeit für die Durchführung des Tests inklusive Instruktion auf 45 Minuten begrenzt. Wäre den teilnehmenden Schülern und Lehrkräften eine längere Testzeit zumutbar gewesen, wäre der Teilbereich der numerischen Intelligenz ebenfalls erfasst worden.

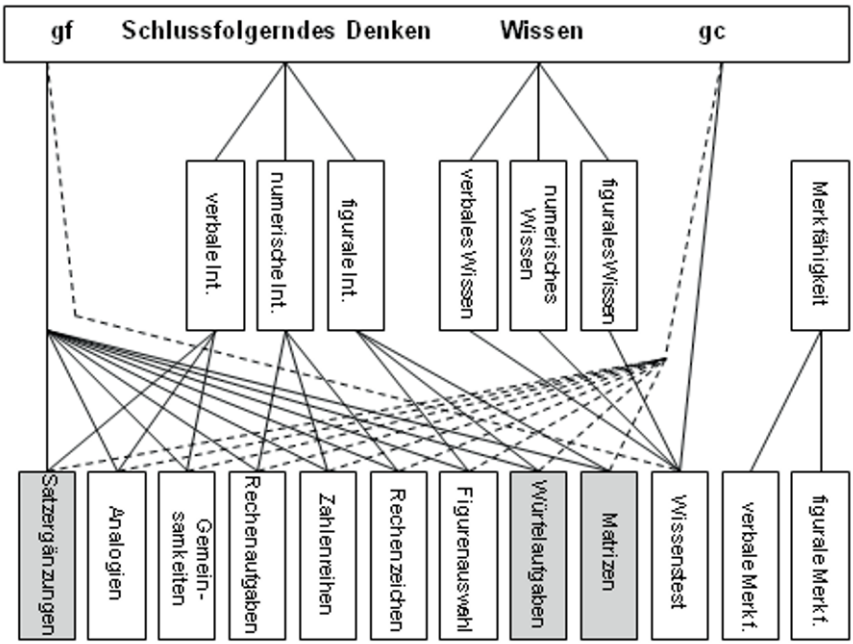


Abbildung 33: Die mit dem I-S-T 2000 R erfasste Fähigkeitsstruktur, (entnommen aus Liepmann et al., 2007, S. 13). In der Studie eingesetzte Subskalen wurden von der Autorin grau hervorgehoben, gf = schlussfolgerndes Denken, gc = Wissen (ebd., S. 12)

Durch die Verwendung einer Auswahl an Subskalen lässt sich zwar keine umfassende Diagnose der Intelligenz auf Individualebene erstellen. Dies ist jedoch für die Umsetzung der Studie auch nicht nötig, da lediglich der Einfluss besonders relevanter kognitiver Fähigkeiten auf die beiden abhängigen Variablen „Wissen und Problemlösen beim Umgang mit (multiplen) Repräsentationen“ sowie auf das „konzeptuelle Verständnis“ statistisch kontrolliert werden sollte.

Tabelle 12 Auswahl an Subskalen zur Erfassung relevanter kognitiver Fähigkeiten

Subskala	Teilbereich	Aufgabenart	Dauer
Satz-ergänzung	verbale Intelligenz	„Jede Aufgabe besteht aus einem Satz, in dem ein Wort fehlt. Aus fünf vorgegebenen Wörtern soll jenes ausgewählt werden, das den Satz richtig vervollständigt“ (zit. n. Liepmann et al., 2007, S. 18).	6 min.
Würfelaufgaben:	figurale Intelligenz (räumlich)	„In der Aufgabe werden Würfel vorgegeben, auf denen jeweils sechs verschiedene Muster abgebildet sind, drei davon sichtbar. Die auszuwählenden Würfel zeigen einen der vorgegebenen Würfel in veränderter Lage. Es soll herausgefunden werden, um welche Würfel es sich jeweils handelt“ (zit. n. ebd., S. 18).	9 min.
Matrizen	figurale Intelligenz (logisch)	„Es werden Anordnungen von Figuren vorgegeben, die nach einer bestimmten Regel aufgebaut sind. Aus vorgegebenen Auswahlfiguren soll jeweils die regelkonforme Lösung herausgefunden werden“ (zit. n. ebd., S. 19).	10 min.

Vor jedem Aufgabentyp erhielten die Schüler eine kurze Instruktion anhand einer Overheadfolie, auf welcher an zwei Beispielen die Aufgabe erklärt wurde (vgl. ebd., im Testheft, S. 7, S. 17, S. 19). Parallel dazu konnten sie die Instruktion im Testheft mitlesen. Das erste Aufgabenbeispiel wurde von der Versuchsleiterin erläutert, das zweite wurde von einem der Schüler erklärt. Der Test startete gemeinsam, wenn die Teilnehmer keine Rückfragen zur Aufgabenstellung mehr stellten und signalisierten, dass sie die Fragestellung verstanden hatten. Die Durchführung des Tests benötigte inklusive Instruktion eine volle Schulstunde (45 Minuten).

Jede Aufgabengruppe umfasst 20 Fragen. Pro Aufgabe standen jeweils fünf Antwortmöglichkeiten zur Auswahl, aus denen die Schüler eine korrekte Antwort wählen sollten. Pro Aufgabe konnte ein Punkt erzielt werden, je Aufgabengruppe also maximal 20 Punkte. Die jeweils erzielten Punktzahl pro Aufgabengruppe wurden entsprechend der individuell erzielten Gesamtpunktzahl in der Aufgabengruppe dem jeweiligen Normwert der Altersgruppe (vgl. ebd., S. 96) zugeordnet. Die teilnehmenden Schüler durften keine Hilfsmittel verwenden.

Des Weiteren wurde vor Beginn der Unterrichtsreihe von allen teilnehmenden Schülern die Vorleistung von den Lehrern erfragt; verwendet wurden die jeweils aktuellsten Zeugnisnoten in den Fächern, Mathematik, Deutsch und Physik. In

Klassen, die zuvor keinen Physikunterricht erhalten hatten, wurde ersatzweise anstelle der Physiknote die Note im Fach Naturwissenschaften (aus Klassenstufe 6) erfasst.

Als weitere Einflussfaktoren wurden zudem die Klassengröße, die Geschlechtszugehörigkeit und der Schultyp (Gymnasium versus Integrierte Gesamtschule) berücksichtigt.

2.3.4.6 Einflussfaktoren bei der Anwendung des Lehrmaterials

Um Einblick in die Umsetzung der Studie zu erlangen und zu sehen, wie die teilnehmenden Lehrkräfte und Schüler mit den zur Verfügung gestellten Materialien umgehen, wurden die Unterrichtsstunden so oft wie möglich vor Ort beobachtet.

Des Weiteren wurden die Schüler im Rahmen des Motivationsfragebogens zu allen drei Messzeitpunkten um eine Einschätzung des Engagements ihres Lehrers gebeten (siehe Erläuterungen zum Motivationsfragebogen).

Zudem wurden die Lehrer in einer offenen Frage gebeten, in einem zur Verfügung gestellten Notizheft besondere Vorkommnisse im Unterricht in eigenen Worten zu notieren, was zumindest einen groben Einblick in die Umsetzung der Stunden erlaubt, in denen aus organisatorischen Gründen (z.B. zeitliche Überschneidung von Hospitation und Datenerhebung an unterschiedlichen Schulen) keine Hospitation möglich war. Die hier verwendeten Items basierten auf Items zur Erfassung von Unterrichtsqualität von Helmke et al. (2010). Pro Stunde wurden acht kurze Fragen zu folgenden Themen gestellt (siehe Anhang C7 auf der Produktseite zu diesem Buch unter www.springer.com):

Einschätzung der Aufgabenschwierigkeit (Bewertung auf einer Skala von 0 zu einfach bis 3 zu schwierig)

- Einschätzung des Aufgabenumfangs (Bewertung auf einer Skala von 0 zu wenige Aufgaben bis 3 zu umfangreich)
- Fragen zum Unterrichtsablauf (Zustimmung zu drei Aussagen, Skala von 1 stimme nicht zu bis 4 stimme zu)
- Bewertung der Lernbilanz (Zustimmung zu drei Aussagen, Skala von 1 stimme nicht zu bis 4 stimme zu)

2.3.5 Ergebnisse zur Messung der abhängigen Variablen

2.3.5.1 Itemstatistiken zum Leistungstest

Um zu untersuchen, wie sich die Änderungen der Aufgaben im Vergleich zur Pilotstudie statistisch auswirken, wurde zunächst eine Itemanalyse des Leistungstests durchgeführt. Eine Übersicht über die Itemmittelwerte und Standardabweichungen findet sich in Tabelle 8 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com.

Die Analyse der Itemschwierigkeit ergab, dass für den Posttest alle Items mit Ausnahme von Item 2b und Item 6 innerhalb des Toleranzbereichs von $0.20 \leq P_i \leq 0.80$ liegen (vgl. Abbildung 34*:* Itemschwierigkeiten des Leistungstests prä, post und follow-up, 2Z: Aufgabenteil Zeichnung, 4b - E: Aufgabenteil Erklärung, 5c-E1: Aufgabenteil Erklärung 1, 5 c-E2: Aufgabenteil-Erklärung 2).

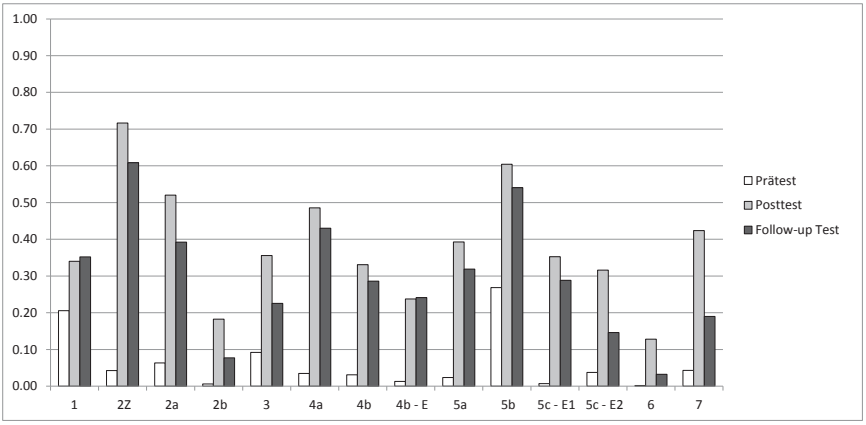


Abbildung 34: Itemschwierigkeiten des Leistungstests prä, post und follow-up, 2Z: Aufgabenteil Zeichnung, 4b - E: Aufgabenteil Erklärung, 5c-E1: Aufgabenteil Erklärung 1, 5 c-E2: Aufgabenteil-Erklärung 2

Für den Prätest zeigen sich durchweg deutlich geringere Werte. In der Follow-up Messung fällt die Lösungswahrscheinlichkeit der Items 6 und 2b Items unter die kritische Grenze von 20%. Zudem sinkt die Lösungswahrscheinlichkeit für

Item 5c (Teil 2)[25] auf einen Wert von 15% (vgl. Tabelle 9 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com). Bemerkenswert ist, dass die Lösungswahrscheinlichkeit für die Wahl der korrekten Gegenstandsweite (5c - Erklärung1) deutlich weniger stark sinkt als die Lösungswahrscheinlichkeit für die Erklärung dieser Wahl (5c - Erklärung2), siehe auch Abbildung 35.

In der Analyse der Trennschärfe zeigt sich, dass die Werte aller Items in der Post -Messung in einem Bereich von $0.29 \leq r_{it} \leq 0.56$ und in der Follow-up Messung in einem Bereich von $0.18 \leq r_{it} \leq 0.59$ liegen (vgl. Abbildung 35). Negative Werte in der Trennschärfe bestanden lediglich in Aufgabe 7 (Anwendung der Abbildungsgleichung) im Vortest (vgl. Tabelle 9 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

Item 6 zeichnet sich sowohl in der Post- als auch in der Follow-up Messung durch die geringste Trennschärfe aus. Die Analyse der Verteilung des Items zeigt, dass das Item rechtssteil und linksschief verteilt ist. Konsistent mit diesem Befund ist die geringe Lösungswahrscheinlichkeit des Items. Daher kann in Erwägung gezogen werden Item 6 auszuschließen.

Cronbachs Alpha als Maß für die interne Konsistenz und Schätzer für die Reliabilität kann mit $\alpha = .77$ im Post- und $\alpha = .75$ im Follow-up Test als zufriedenstellend und mit $\alpha = .29$ als gering im Prätest bewertet werden. Unter Ausschluss von Item 6 ergeben sich die gleichen Werte. Der niedrige Wert im Vortest ist auf das geringe Vorwissen der Schüler zurückzuführen, die zuvor keinen Unterricht in Strahlenoptik erhalten hatten. Die Test-Retest-Reliabilität liegt mit $r_{\text{Test-Retest}}(482) = .44$, $p < .01$ im akzeptablen Bereich. Da die zeitliche Stablität des Merkmals aufgrund des geringen Vorwissens der Schüler nicht vorausgesetzt werden kann, wurde davon abgesehen die Retest-Reliabilität prä – post und prä – follow-up zu berechnen.

25 Wortlaut des Items: (Erklärung – Teil 2: Kreuze an, wie groß die Gegenstandsweite *g* etwa sein muss! Erkläre, wie du auf die Lösung des Aufgabenteils b) gekommen bist!)

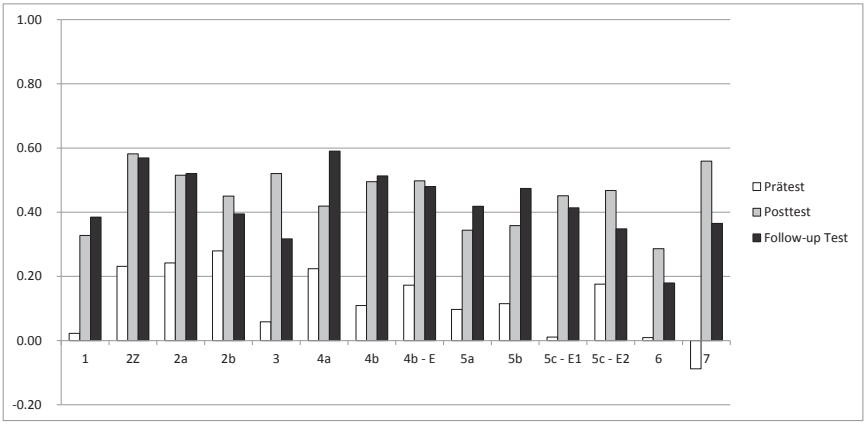


Abbildung 35: Korrigierte Trennschärfen der Items des Leistungstests, prä, post, und follow-up, 2Z: Aufgabenteil Zeichnung, 4b - E: Aufgabenteil Erklärung, 5c - E1: Aufgabenteil Erklärung 1, 5c - E2: Aufgabenteil-Erklärung 2

2.3.5.2 Faktorenanalyse des Leistungstests

Um die Dimensionalität des Instruments (die Struktur des Tests) zu analysieren, wurde eine konfirmatorische Faktorenanalyse durchgeführt. Ergebnisse einer exploratorischen Faktorenanalyse des Instruments auf Basis einer von den hier berichteten Daten unabhängigen Stichprobe, finden sich in der Dissertation von Scheid (2013).

Mittels der konfirmatorischen Faktorenanlyse wurde zunächst untersucht, ob es sich um ein eindimensionales Testinstrument handelt. Hierzu wurde für jeden der drei Messzeitpunkte getrennt ein Generalfaktorenmodell aufgestellt, bei dem alle Variablen auf einen Faktor laden. In diesem Modell wird angenommen, dass alle 13 Variablen nur ein einziges latentes Merkmal (hier das physikalische Verständnis beim Umgang mit (multiplen) Repräsentationen) erfassen und darüber hinaus nur zufällige Messfehler und indikatorspezifische Varianz. Im Ergebnis zeigt sich, dass das Generalfaktorenmodell für jeden der drei Messzeitpunkte verworfen werden muss, da das Modell außerhalb der in der Literatur angegebenen Grenzen für Cut-off-Werte liegen: *CFI* und *TLI* ≥ 0.95 – guter Bereich bzw. *CFI* ≥ 0.90 – akzeptabler Bereich; *RMSEA* < 0.06 ($N > 250$), *SRMR* < 0.11 (vgl. Bühner, 2011, S. 428). Exemplarisch werden an dieser Stelle die Fit-Indizes des Generalfaktorenmodells für den zweiten Messzeitpunkt aufgeführt, um die Ablehnung des Generalfaktorenmodells zu belegen ($\chi^2_{(65)} = 308.35, p < .001$; $CFI = 0.83$; $TLI =$

0.80; *RMSEA* = 0.09; *SRMR* = 0.06). Die Fit-Indizes des dritten Messzeitpunktes weichen hierbei nur geringfügig von den Ergebnissen des zweiten Messzeitpunktes ab. Für den ersten Messzeitpunkt ergeben sich die deutlichsten Abweichungen von den Untergrenzen der Cut-Off-Werte. Offenbar handelt es sich um ein mehrdimensionales Testinstrument, das unterschiedliche Fähigkeitsfacetten oder möglicherweise sogar voneinander unabhängige Fähigkeiten erfasst.

Untersucht wurde hierbei zum einen die inhaltliche Struktur des Tests, d.h. Klärung der Frage, welche Items inhaltlich zusammenhängen und ein latentes Merkmal abbilden und zum anderen die Frage, wie stark die verschiedenen Facetten inhaltlich miteinander verbunden sind. Bei der Durchführung der Analyse zeigte sich, dass sich die Interpretierbarkeit der Modelle unter Ausschluss von Item 13 (Aufgabe 6) deutlich erhöhte, welches bereits bei der Analyse der Itemstatistiken aufgefallen war. Zur Analyse der inhaltlichen Struktur wurde ein 2-Faktorenmodell einem 3-Faktorenmodell gegenübergestellt (siehe Tabelle 14).

1. Im 2-Faktorenmodell wurde folgende Unterscheidung getroffen: Ein Faktor, welcher das konzeptuelle Verständnis im Umgang mit Repräsentationen und ein Faktor, welcher nur den Umgang mit Repräsentationen (ohne Thematisierung von Schülervorstellungen) erfasst.
   - Faktor 1 („Schülervorstellungen“) = y_1, y_6, y_7, y_8
   - Faktor 2 („Umgang mit Repräsentationen)“ = y_2, y_3, y_4, y_5, y_9, y_{10}, y_{11}, y_{12}, y_{14}
2. Im 3-Faktorenmodell wurden drei Faktoren angenommen, welche auf den Umgang mit verschiedenen Repräsentationsformen zielen: Umgang mit deskriptiven Repräsentationen, Umgang mit depiktional realistischen Repräsentationen, die sich eng am Experiment orientieren und Umgang mit depiktional schematischen Repräsentationen.
   - Faktor 1 („Umgang mit deskriptiven Repräsentationen“) = y_1, y_3, y_4, y_5, y_8, y_{14}[26]. Der erste Faktor bezieht sich auf das Ziehen logischer Schlüsse auf Basis des Operierens mit deskriptiven Repräsentationen. Die deskriptiven Repräsentationen schließen hierbei sowohl verbale Begründungen als auch die Verwendung der Abbildungsgleichung (mathematische Repräsentation) ein.
   - Faktor 2 („Bezüge zu einem realen Experiment herstellen)“= y_9, y_{10}, y_{11}, y_{12}.[27]. Der zweite Faktor bezieht sich ausschließlich auf alle Teilfragen

26 entspricht den Aufgaben: 1, 2a, 2b, 3, 4b - E (4b - Aufgabenteil Erklärung), 7

27 entspricht den Aufgaben: 5a, 5b, 5c - E1 (Aufgabenteil c, Erklärung 1), 5c - E2 (Aufgabenteil c, Erklärung 2)

von Aufgabe 5, in der es darum ging, aus einer depiktional realistischen Repräsentation, in der ein reales Experiment dargestellt ist, Informationen abzulesen.

- Faktor 3 („Umgang mit schematisch depiktionalen Repräsentationen“) = y_2, y_6, y_7[28]. Der dritte Faktor bezieht sich auf das eigene Erstellen von schematisch deskriptiven Repräsentationen der Strahlenkonstruktion und dem Operieren mit denselben.

Um zu analysieren, wie stark die verschiedenen Facetten miteinander zusammenhängen wurde jedes der beiden Modelle (2-Faktorenmodell versus 3-Faktorenmodell), nun auf unterschiedliche Weise spezifiziert:

- Basismodell mit Korrelationen: Zunächst wurde jeweils ein Modell aufgestellt, bei dem die jeweiligen Items auf zwei bzw. drei Faktoren laden und das Programm die Korrelationen zwischen den Faktoren schätzt.
- Bifaktormodell: In einem nächsten Schritt wurde ein bifaktorielles Modell aufgestellt. Gemäß diesem Modell bestehen ebenso wie im Basismodell zwei bzw. drei Faktoren, die für die Erfassung voneinander unabhängiger Merkmale stehen. Korrelationen zwischen den Faktoren wurden daher nicht zugelassen. Über diese zwei bzw. drei Faktoren hinaus wird angenommen, dass ein Generalfaktor existiert, welcher die Bearbeitung aller Items beeinflusst.
- Second-Order-Modell: Im letzten Schritt wurde ein Second-Order-Modell aufgestellt. In diesem Modell wird angenommen, dass ein Faktor erster Ordnung besteht, der sich auf zwei bzw. drei Faktoren zweiter Ordnung auswirkt, welche für voneinander unabhängige latente Merkmale stehen. Entsprechend sind die Faktoren zweiter Ordnung untereinander unkorreliert.
- Insgesamt wurden zusätzlich zum Generalfaktorenmodell also sechs Modelle zu drei Zeitpunkten verglichen. Die Berechnung der Modelle erfolgte unter Verwendung der freien Statistiksoftware R und der Pakete lavaan (Rosseel, 2012), qgraph (Epskamp et al., 2012) und psych (Revelle, 2013).

Im Hinblick auf die Fit-Indizes zeigt sich, dass von den getesteten Modellen das Bifaktormodell „Umgang mit unterschiedlichen Repräsentationsformen“ innerhalb der in der Literatur angegebenen Grenzen der Cut-off-Werte liegt (vgl. Tabelle 13 und Tabelle 14). Das Modell (vgl. Abbildung 37) kann inhaltlich wie folgt interpretiert werden: Es existiert eine generelle Kompetenz, welche die Be-

28 entspricht den Aufgaben: 2Z (Zeichnung), 4a und 4b

arbeitung aller gestellten Aufgaben beeinflusst. Dass diesbezüglich jeder der Pfadkoeffizienten signifikant wurde, stützt die Gültigkeit dieser Annahme. Des Weiteren existieren drei Subskalen, die über diese Fähigkeit hinaus den „Umgang mit deskriptiven Repräsentationen", die Fähigkeit Bezüge zu einem realen Experiment herzustellen" und „den Umgang mit schematisch depiktionalen Repräsentationen" erfassen (vgl. Abbildung 36). Jede dieser Kompetenzen ist voneinander unabhängig.

Tabelle 13 Fit-Indizes: Konfirmatorische Faktorenanalysen Leistungsposttest, ($N = 484$)

	2-Faktorenmodell (inklusive Skala zum Konzeptuellen Verständnis)			3-Faktorenmodell (Umgang mit unterschiedlichen Repräsentations-formen)		
	[a]BMK	[b]BFM	[c]SOM	[a]BMK	[b]BFM	[c]SOM
$\chi^2_{(df)}$	257.47 $_{(64)}^{***}$	152.95 $_{(52)}^{***}$	257.47 $_{(63)}^{***}$	177.72 $_{(62)}^{***}$	112.78 $_{(52)}^{***}$	169.85 $_{(61)}^{***}$
CFI	0.87	0.93	0.87	0.92	0.96	0.92
TLI	0.84	0.90	0.83	0.90	0.94	0.89
RMSEA	0.08	0.06	0.08	0.06	0.05	0.06
SRMR	0.06	0.04	0.06	0.05	0.04	0.05
AIC	14184.38	14103.86	14186.38	14108.63	14063.69	14112.77
BIC	14297.30	14266.96	14303.48	14229.91	14226.79	14238.23

*** $p < .001$

[a]BMK: Basismodell mit Korrelationen zwischen Faktoren, [b]BFM: Bifaktormodell, [c]SOM: 2nd Order Modell

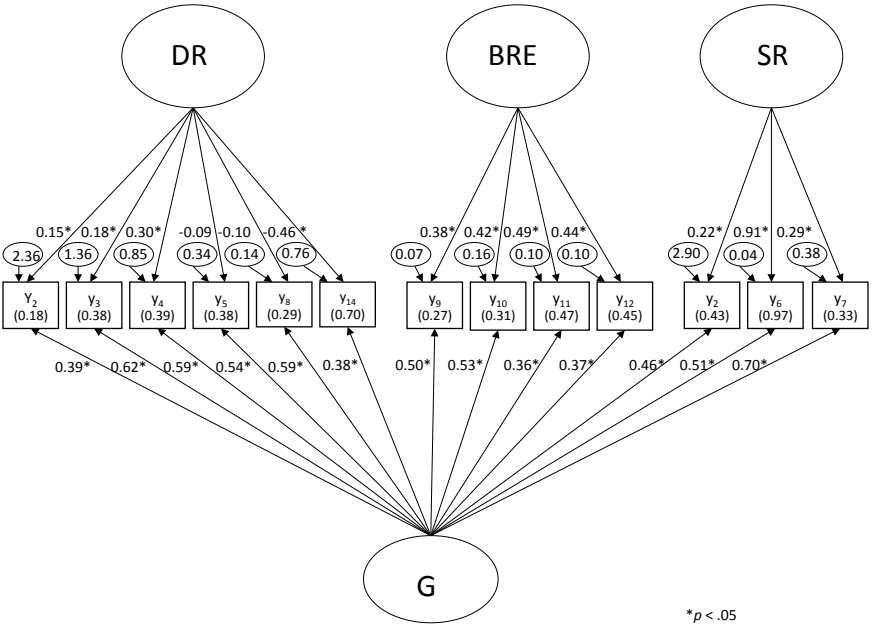


Abbildung 36: Bifaktormodell „Umgang mit unterschiedlichen Repräsentationsformen“, DR = Umgang mit deskriptiven Repräsentationen, BRE = Bezüge zu einem realen Experiment herstellen, SR = Umgang mit schematisch depiktionalen Repräsentationen, G = Generalfaktor

Da jedoch nicht alle Pfadkoeffizienten für die Pfade zwischen den Items und denjenigen Faktoren, welche die Subskalen abbilden, signifikant sind, ist auch dieses Modell mit Vorsicht zu interpretieren. Dies betrifft die Items: y_5 und y_8. Bei diesen Items wurden die betreffenden Pfadkoeffizienten nicht signifikant. Geringere, noch akzeptable globale Fitwerte ($\chi^2_{(62)} = 177.72, p < .001$; *CFI* = 0.92; *RMSEA* = 0.06; *SRMR* = 0.05), jedoch bessere lokale Fitwerte, weist das Basismodell mit Korrelationen zwischen den drei Faktoren auf (vgl. Tabelle 13).

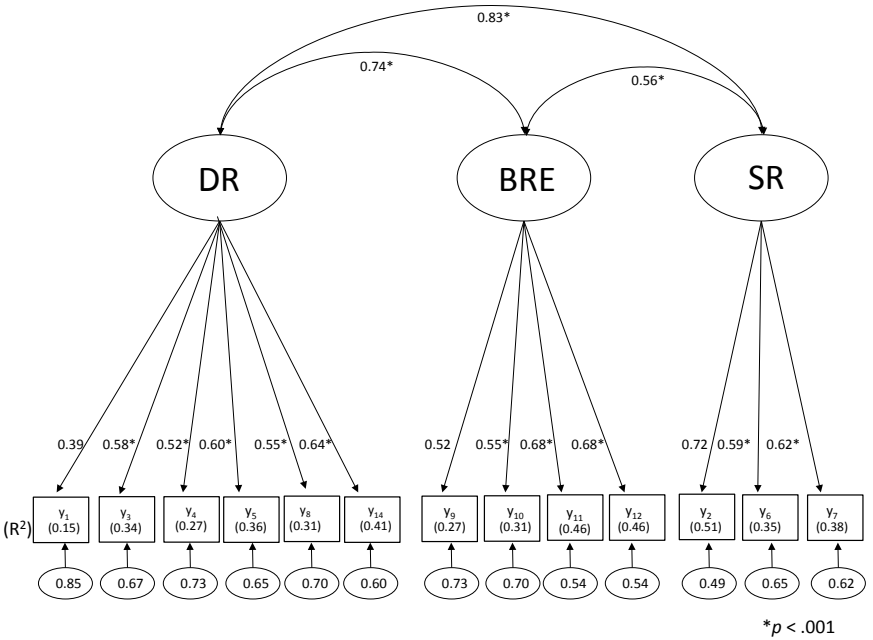


Abbildung 37: BMK: Basismodell mit Korrelationen zwischen Faktoren „Umgang mit unterschiedlichen Repräsentationsformen", DR = Umgang mit deskriptiven Repräsentationen, BRE = Bezüge zu einem realen Experiment herstellen, SR = Umgang mit Umgang mit schematisch depiktionalen Repräsentationen

Zusammenfassend lässt sich festhalten, dass sich die Physikleistung durch drei Kernkompetenzen beschreiben lässt, die vermutlich jedoch durch eine generelle „Repräsentations"-Kompetenz beeinflusst werden.

Tabelle 14 Zuordnung der Items des Leistungstests zu den Skalen „Umgang mit verschiedenen Repräsentationsformen, Reliabilität der Skalen und lokale Gütemaße, ($N = 484$)

F	Item / Aufgabe		Indikatorreliabilität	Faktor-reliabilität	*DEV*	Formel Larcker Kriterium	*α*
DR[a]	y_1	1	0.15	.68	0.29	.69 (nicht erfüllt)	.66
	y_3	2a	0.34				
	y_4	2b	0.27				
	y_5	3	0.36				
	y_8	4b -E	0.30				
	y_{14}	7	0.41				
BRE[b]	y_9	5a	0.27	.70	0.38	>.54	.69
	Y_{10}	5b	0.31				
	Y_{11}	5c -E1	0.46				
	Y_{12}	5c -E2	0.46				
SR[c]	Y_2	2Z	0.15	.67	0.47	>.31	.57
	Y_6	4a	0.31				
	Y_7	4b	0.16				

Erläuterungen

y_i		Item	Faktor
y_1	1	**Deskriptive R. / Verbale Begründung/ Schülervorstellung:** Tobias möchte an einem sonnigen Tag mit Hilfe einer Sammellinse ein Streichholz entzünden. Nadine schlägt Tobias vor, eine verstellbare Lochblende vor die Sammellinse zu setzen. Kann Tobias dadurch das Streichholz besser entzünden?	**DR**[a]
y_3	2a	**Deskriptive R. / Verbale Begründung:** Wie würde sich die Bildgröße und Bildweite verändern, wenn man den Gegenstand weiter von der Linse entfernt?	
y_4	2b	**Deskriptive R. / Verbale Begründung:** Begründe mit den Strahlen in der Abbildung oben: Warum verändert sich die Bildgröße so wie oben beschrieben, wenn der Gegenstand weiter von der Linse entfernt wird?	
y_5	3	**Deskriptive R. / Deskriptive Darstellung:** Ein Gegenstand befindet sich vor einer Linse. Die Gegenstandsweite liegt zwischen einfacher und doppelter Brennweite der Linse. Wie groß ist das Bild verglichen mit der Gegenstandsgröße? Wie groß ist die Bildweite verglichen mit der Brennweite?	
y_8	4b - E	4b Aufgabenteil Erklärung **Schülervorstellung/Lochblende/Strahlenkonstruktion/verbale Beschreibung:** Erkläre in Worten, was mit dem Bild passiert.	
y_{14}	7	**Deskriptive R. / Berechnung:** Berechne *B* bei einer scharfen Abbildung eines Gegenstandes mit einer Sammellinse.	
y_9	5a	**Bezug reales Experiment / Fotografie** Ein gewöhnliches Zimmerfenster wird durch eine Sammellinse auf einer Wand in einem Raum abgebildet. Um welchen Bildfall handelt es sich?	**BRE**[b]
y_{10}	5b	**Bezug reales Experiment / Fotografie** Kreuze an, wie groß die Gegenstandsweite g etwa sein muss!	
y_{11}	5c - E1	Aufgabenteil Erklärung 1 **Bezug reales Experiment / Fotografie:** Erkläre, wie du auf die Lösung des Aufgabenteils a) gekommen bist!	

y_{12}	5c - E2	Aufgabenteil Erklärung 2 **Bezug reales Experiment / Fotografie**: Erkläre, wie du auf die Lösung des Aufgabenteils b) gekommen bist!	
y_2	2Z	Aufgabe 2 Zeichnung **Schematisch depiktionale R. / Strahlenkonstruktion:** Zeichne den Strahlengang des folgenden Versuchsaufbaus.	**SR**[c]
y_6	4a	**Schematisch depiktionale R. / Problemlöseaufgabe / Strahlenkonstruktion**: Abbildung (unten) zeigt einen Gegenstand (hier als Pfeil dargestellt) mit der Gegenstandsgröße G und sein Bild, das durch die Sammellinse entsteht, mit der Bildgröße B. Der Bildpunkt Q zum Gegenstandspunkt P wurde richtig konstruiert. Die Strahlenkonstruktion ist nicht dargestellt. Der Schirm ist jedoch nicht an der richtigen Stelle aufgestellt, sondern ein Stück zur Linse hingerückt. a) Zeige durch Einzeichnen, dass auf dem Schirm anstatt des Bildpunktes Q ein unscharfer Bildfleck entsteht. Verwende hierfür für die gegebene Abbildung (unten).	
y_7	4b	**Schematisch depiktionale R. / Schülervorstellung / Lochblende / Strahlenkonstruktion:** Wie ändert sich das Bild des Gegenstandes, wenn man eine Lochblende vor die Linse stellt? Begründe Deine Antwort durch eine Zeichnung. Verwende hierfür die gegebene Abbildung (unten).	

[a] DR = Umgang mit deskriptiven Repräsentationen
[b] BRE = Bezüge zu einem realen Experiment herstellen
[c] SR = Umgang mit Umgang mit schematisch depiktionalen Repräsentationen

2.3.5.3 Itemstatistiken zum Konzepttest

Zur Analyse, wie sich die Überarbeitung der Aufgaben statistisch auswirkt, wurde für den Konzepttest ebenfalls eine Itemanalyse durchgeführt (vgl. Tabellen 11-13 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

Die zugrunde liegende Stichprobe basiert neben den Daten der hier diskutierten Studie auf den Daten von Jochen Scheid. Bei den Daten von Jochen Scheid handelt es sich um eine Stichprobe von $N = 444$ Schülern, die ebenfalls der 7. und 8. Klassenstufe angehörten, die allerdings ausschließlich das Gymnasium besuchen. Detailliertere Angaben zur Stichprobe finden sich in der Dissertationsschrift von Jochen Scheid (vgl. Scheid, 2013). Ziel der Studie von Jochen Scheid war es, Effekte der Förderung der Kohärenz im Umgang mit Repräsentationen auf den Lernzuwachs zu untersuchen. In der Studie von Jochen Scheid wurde der Konzepttest ebenfalls zu drei Messzeitpunkten verwendet: vor der Intervention, nach der Intervention, welche in dieser Studie einen vergleichbaren Umfang von sechs Stunden hatte, sowie sechs Wochen später. Durchführung und Auswertung des Tests erfolgten in gleicher Weise wie in der hier dargestellten Studie. Die Testfassung war exakt identisch.

In einem ersten Schritt wurden die Mittelwerte der beiden Stichproben zum ersten Messzeitpunkt verglichen, als die Schüler in beiden Stichproben noch keinen Unterricht zur Bildentstehung bei der Sammellinse erhalten hatten.

Die Stichproben werden wie folgt unterschieden:

- Die Stichprobe von Jochen Scheid wird im Folgenden mit Stichprobe „Kohärenz“ abgekürzt „Ko“ bezeichnet, da Schüler in dieser Studie in der Treatmentgruppe Unterricht zur Förderung der Kohärenz von Repräsentationen erhielten, während Schüler der Kontrollgruppe herkömmliche Aufgaben bearbeiteten.
- Die Stichprobe, welche der hier vorgestellten Studie zu Grunde liegt, wird Stichprobe „Schülervorstellungen“ genannt und im Folgenden mit „SV“ abgekürzt, da der Unterricht in beiden Bedingungen (Treatment- und Kontrollgruppe) auf die Förderung des konzeptuellen Verständnisses und damit einhergehend darauf zielte, Schülervorstellungen zu überwinden.

Da in der Studie zur Förderung der Kohärenz kein Unterschied zwischen Treatment- und Kontrollgruppe gefunden wurde (vgl. Scheid, 2013), wurde darauf verzichtet, die Itemstatistiken separat je Bedingung (Treatment- versus Kontrollgruppe) zu berichten. Detailliertere Angaben zu den Ergebnissen des Konzepttests für die Stichprobe „Ko“ finden sich in Scheid (2013, S. 135 ff.). Analog dazu wurden auch in der Studie zu Schülervorstellungen so verfahren, dass Treatment- und Kontrollgruppe gesamt betrachtet wurden. Der Vergleich zielt also darauf, Unterschiede zwischen den Stichproben bzw. den Studien auf Itemebene darzustellen, welche darauf beruhen, dass in der Studie „SV“ explizit auf die Förderung des konzeptuellen Verständnisses eingegangen wurde, während in der Studie „Ko“ andere Lernziele im Vordergrund standen.

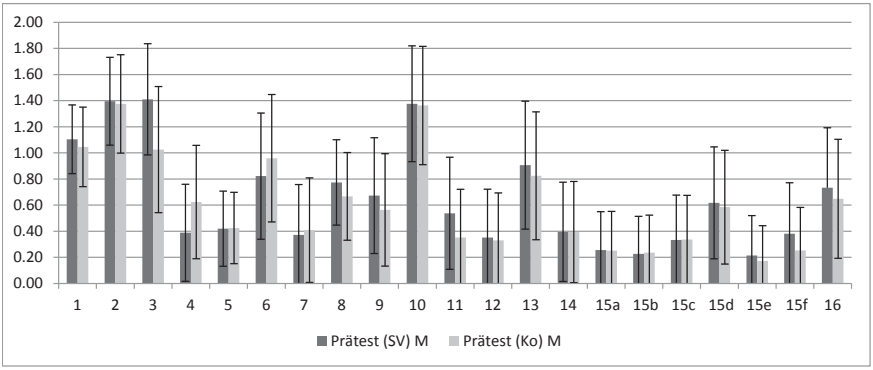


Abbildung 38: Mittelwerte und Standardabweichungen des Konzeptprätest der Stichproben „SV“ (= Schülervorstellungen) und „Ko“ (= Kohärenz) im Vergleich (nur Gymnasiasten)

Im ersten Schritt wurde analysiert, ob sich die Stichproben „SV“ und „Ko“ bereits zu Beginn systematisch unterscheiden, was zum ersten Messzeitpunkt erwartungsgemäß nicht der Fall sein sollte. Hierzu wurden in der Stichprobe „SV“ ausschließlich Gymnasiasten einbezogen, um größtmögliche Vergleichbarkeit zwischen den Stichproben zu erreichen (vgl. Tabelle 10 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com und Abbildung 38). Bezüglich der Gesamtpunktzahl im Konzeptprätest lassen sich keine Unterschiede zwischen den Gruppen feststellen ($t(860) = .64$, *n.s.*).

In den folgenden Itemstatistiken wurden die Daten aller Schüler der Stichproben „SV“ und „Ko“ verglichen, also auch die Daten der Gesamtschüler in der Stichprobe „SV“ einbezogen. Erwartungsgemäß erzielten die teilnehmenden Schüler nach dem Unterricht zur Bildentstehung bei der Sammellinse im Schnitt durchweg höhere Punktzahlen als zum ersten Messzeitpunkt, unabhängig davon, ob gezielt auf die Förderung des konzeptuellen Verständnisses eingegangen wurde oder nicht (siehe Tabelle 11 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

Die Analyse der Itemschwierigkeit ergab, dass für den Posttest der Stichprobe „SV“ alle Items mit Ausnahme von Item 15e innerhalb des Toleranzbereichs von $0.20 \leq P_i \leq 0.80$ liegen (vgl. auch Abbildung 39). Die Lösungswahrscheinlichkeit beträgt für Item 15e 15 % (vgl. Tabelle 12 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

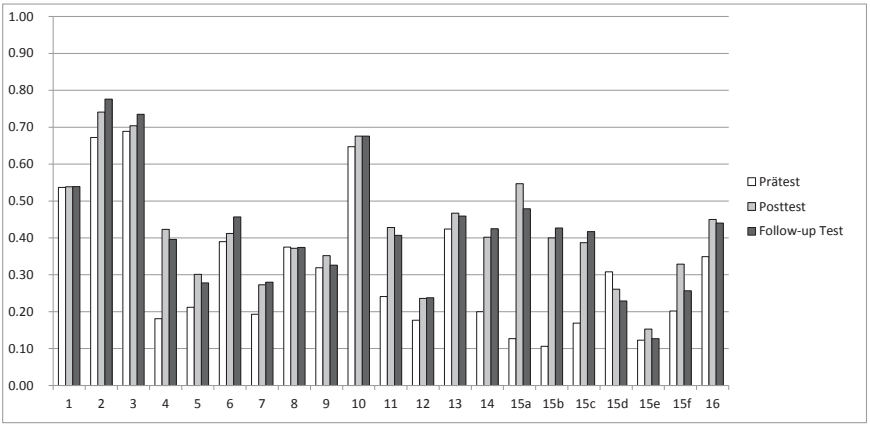


Abbildung 39: Abbildung: Itemschwierigkeiten des Konzepttests für die Stichprobe „SV“ (= Schülervorstellungen) je Messzeitpunkt

In Bezug auf die Stichprobe „SV“ bestätigt sich beim Blick auf die Lösungswahrscheinlichkeiten, dass sich der Unterricht stärker auf Bearbeitung von Aufgaben auswirkt, die spezielle Konzepte zur Bildentstehung thematisieren, wie etwa den Strahlengang, die Bildentstehung unter Verwendung einer Blende oder Fragen zum virtuellen Bild (Item 11 bis 16), als auf die Bearbeitung von Aufgaben, in denen es allgemein um Konzepte der Lichtausbreitung, gerichteten Reflexion, Streuung und der physikalischen Sehvorstellung geht (Items 1-10), siehe auch Abbildung 39. Item 7 und 15a fallen zudem ebenfalls mit einer geringen Lösungswahrscheinlichkeit im Posttest und Follow-up Test der Stichprobe „Ko“ auf (vgl. Tabelle 13 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

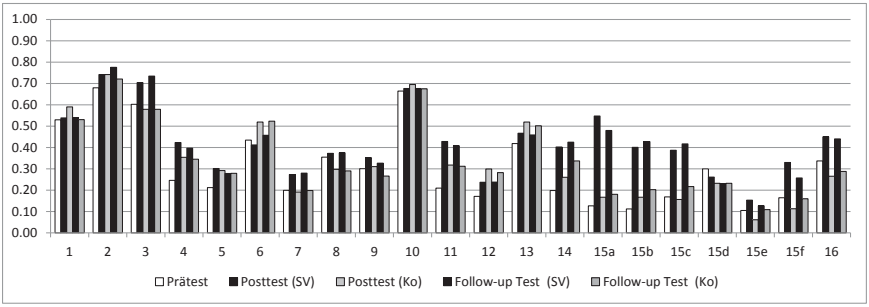


Abbildung 40: Itemschwierigkeit für den Konzeptprätest: Prätest: Gesamt, Post- und Follow-up Test je Stichprobe („SV“ = Schülervorstellungen versus „Ko“ = Kohärenz)

In der Stichprobe SV weist Item 15a jedoch „gute“ Werte der Lösungswahrscheinlichkeit zum zweiten und dritten Messzeitpunkt auf, was darauf hindeutet, dass dieses Item sich dafür eignet das Wissen von Schülern zu differenzieren, nachdem die Schüler Unterricht zur Förderung des konzeptuellen Verständnisses erhalten hatten (vgl. Abbildung 40 sowie Tabelle 12 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

Eine abschließende Beurteilung kann hier ein Blick auf die Trennschärfe geben. Die Analyse der Trennschärfe zeigt, dass die Items 7, 15d, 15e eine geringe Trennschärfe aufweisen. Die Items 7 und 15e waren bereits zuvor negativ aufgefallen (vgl. Abbildung 41 sowie Tabelle 13 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).

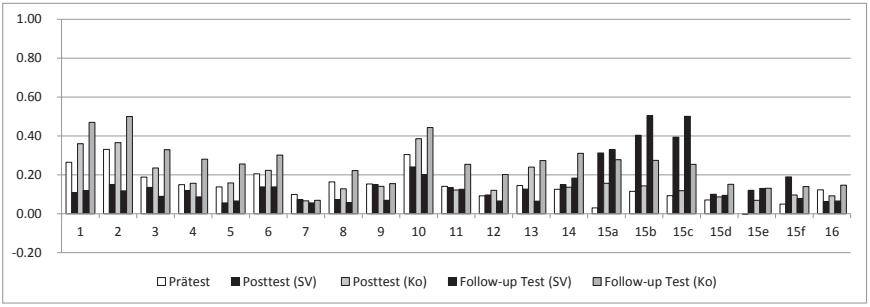


Abbildung 41: Korrgierte Trennschärfe für den Konzeptprätest: Prätest: Gesamt, Post- und Follow-up Test je Stichprobe „SV“ (Schülervorstellungen) versus „Ko“ (Kohärenz)

Des Weiteren geht die Trennschärfe bei einigen Items im Vergleich der Gesamtstichprobe („SV“ und „Ko“) und der Postmessung („SV“) zurück (vgl. Tabelle 13 in Anhang C9). Um diesen Rückgang genauer zu analysieren, wurde die Trennschärfe zu allen drei Messzeitpunkten gesondert nochmals für die Stichprobe „SV“ betrachtet (vgl. Abbildung 42).

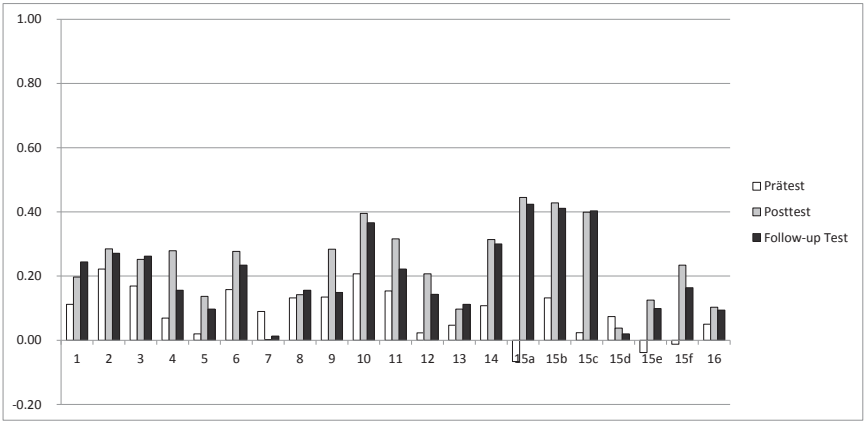


Abbildung 42: korrigierte Trennschärfen des Konzepttests für die Stichprobe „SV“ (Schülervorstellungen) je Messzeitpunkt

Ein Blick auf Abbildung 42 zeigt, dass der Effekt des Rückgangs von Messzeitpunkt 1 zu Messzeitpunkt 2, mit Ausnahme von Item 15d, verschwindet, wenn lediglich die Stichprobe „SV“ zu Grunde gelegt wird.

Die Analyse der Trennschärfe in der Stichprobe „SV“ ergibt zudem, dass die Werte in der Post-Messung zwischen $0.10 \leq r_{it} \leq 0.45$ und in der Follow-up Messung in einem Bereich von $0.10 \leq r_{it} \leq 0.40$ liegen (Tabelle 13 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com). Negative Werte in der Trennschärfe bestanden lediglich im Vortest und zwar in den Items: 15a, 15e und 15f.

Diese negativen Werte liegen jedoch nahe am Bereich 0, was darauf hinweist, dass diese Items spezifisches Wissen zur Bildentstehung bei der Sammellinse erfassen und daher wenig dazu geeignet sind das konzeptuelle Verständnis von Schülern ohne Vorkenntnisse dieser Lerninhalte einzuschätzen. Im Posttest erweisen sie sich jedoch als geeignet.

Zwischen-Fazit: Ein Vergleich der Lösungswahrscheinlichkeiten und der Trennschärfe zwischen den Stichproben von Jochen Scheid (Förderung der repräsentationalen Kohärenz ohne gezielte Berücksichtigung von Schülervorstellungen) und dem Treatment der hier vorgestellten Studie legt nahe, dass die Items 7, 15d und 15e einer Überarbeitung bedürfen. Folgende Schwierigkeiten sind mit den Items verbunden:

- Die ursprüngliche Fassung von Item 7 enthielt eine ungenaue Formulierung, welche zu einer mehrdeutigen Interpretation führte: „In einem großen dunklen Raum stehen Corinna und Nadine nebeneinander vor einem Spiegel. Corinna beleuchtet den Spiegel schräg mit einer Taschenlampe, die ein schmales Lichtbündel erzeugt. Wer kann das Licht der Taschenlampe sehen?“ Je nachdem wie „schmal“ der jeweilige Schüler das Lichtbündel interpretierte, fiel das Licht auf die Beobachterin Nadine oder ging knapp an ihr vorbei.
- Bei Item 15d „Was passiert mit dem Bild, wenn der Schirm entfernt wird?“ und 15e „In einer Versuchsanordnung sind eine Glühlampe, eine Sammellinse und ein Schirm so montiert, dass ein vergrößertes, umgekehrtes, scharfes Bild des Glühfadens entsteht: Wo ist das Bild, nachdem der Schirm weggenommen wurde?“ könnte der Begriff des Bildes problematisch gewesen sein. Möglicherweise wurde der Ort an dem sich die Strahlen gemäß der Modellvorstellung vom Lichtstrahl treffen nicht als „Bild“ interpretiert, da der Beobachter bzw. der Schirm fehlt, weshalb Schüler, denen klar ist, dass sich die Strahlen auch im Raum treffen, diesen Sachverhalt nicht unter dem Begriff „Bild“ subsumieren, sodass zwischen diesen Schülern und Schülern, welche diesen Sachverhalt nicht verstanden haben, nicht ausreichend differenziert werden kann.

In der Diskussion mit einer externen Arbeitsgruppe ergab sich zudem folgendes Problem bezüglich des Item 1 („Du siehst hier auf dem Bild eine brennende Kerze. Wo ist das Licht? Schraffiere den Bereich / die Bereiche, in dem das Licht deiner Meinung nach ist, mit einem Stift“) wurde lediglich mit der vollen Punktzahl bewertet wenn der Schattenraum unterhalb der Kerze ausgespart wurde. Hier sollten die Auswertungskriterien überarbeitet werden, da durch die Rückstreuung auch der Schattenraum erhellt ist, weshalb der Körper der Kerze zu sehen ist.

Aus den genannten Gründen wurden die Items 1, 7 und 15d und 15e von der Gesamtwertung ausgeschlossen. Cronbachs Alpha[29] als Maß für die interne Konsistenz und Schätzer für die Reliabilität verbessert sich unter Ausschluss der Items 1, 7, 15d und 15e systematisch in der Post- und der Follow-up Messung (vgl. Tabelle 15). Der Ausschluss weiterer Items auf Basis der Ergebnisse der Faktorenanalyse, welche in dem folgenden Kapitel vorgestellt wird, führt zu einer weiteren Verbesserung der internen Konsistenz. Insgesamt wurden 11 Items beibehalten, welche in die Errechnung der Gesamtpunktzahl eingingen (maximal erreichbare Höchstpunktzahl: 22 Punkte). Insgesamt kann die interne Konsistenz mit .79 zum zweiten Messzeitpunkt unter Ausschluss der problematischen Items in der Stichprobe „SV“ als gut und in der Gesamtstichprobe mit .75 als zufriedenstellend eingestuft werden. Die Test-Retest Reliabilität kann ebenfalls als zufriedenstellend bewertet werden. Da aufgrund des geringen Vorwissens der Schüler keine zeitliche Stabilität des Merkmals gegeben war, wurde davon abgesehen, die Retest-Reliabilität prä – post und prä – follow-up zu berechnen.

29 Da es sich um ordinalskalierte Daten handelt, wurde Cronbachs Alpha basierend auf einer polychorischen Korrelationsmatrix (vgl. Eid, Gollwitzer & Schmitt, 2010, S. 922) berechnet. Erläuterungen zur Anwendung dieses Verfahrens finden sich in Kapitel 2.2.6.3 Exkurs: Berechnung einer polychorischen Korrelationsmatrix für ordinalskalierte Daten.

Tabelle 15 Reliabilitätsstatistiken für den Konzepttest

	Gesamtstichprobe			Stichprobe „SV“		
	Prätest (N = 935)	Posttest (N = 869)	Follow-up (N = 922)	Prätest (n = 491)	Posttest (n = 480)	Follow-up (n = 486)
α (21 Items)	.37	.69	.64	.42	.75	.71
α (17 Items)[a]	.40	.75	.66	.42	.77	.72
α (11 Items)[b]	.45	.75	.70	.51	.79	.76
$r_{Test\text{-}Retest}$		post – follow-up			post – follow-up	
(21 Items)		.62 ***			.60 ***	
(17 Items)[a]		.62 ***			.62 ***	
(11 Items)[b]		.64 ***			.63 ***	

*** $p < .001$

[a] unter Ausschluss der Items: 1, 7, 15d, 15e

[b] Auf Basis späterer Analysen wurden weitere sechs Items ausgeschlossen (5, 8, 12, 13, 15f, 16) , siehe Kapitel 2.3.5.5 Kreuzvalidierung des Konzepttests.

2.3.5.4 Raschanalyse des Konzepttests

Zur weiteren Analyse des Testinstruments wurde der Test auf Raschskalierbarkeit geprüft. Die zugrundeliegenden Daten basieren auf der Gesamtstichprobe. Hierzu wurde pro Messzeitpunkt ein „Ein-Parameter-Logistisches-Modell für dichotome Items“ aufgestellt. Die Analyse wurde mit dem freien Statistikprogramm R unter Verwendung des Softwarepackage eRm von Poinstingl, Mair und Hatzinger (2007) durchgeführt. Da in den Items 0, 1 oder 2 Punkte erreicht werden konnten, wurden die Items nachträglich wie folgt dichotomisiert:

- Vollständig richtig gelösten Antworten wurde der Wert 1 zugeordnet.
- Antworten, die mit einem Punkt bewertet worden waren, weil die teilnehmende Person sowohl die richtige Antwort als auch einen Distraktor gewählt hatte, wurden mit 0 Punkten bewertet.
- Antworten, die zuvor mit 0 Punkten bewertet wurden, wurden mit 0 Punkten belassen.

Zur Modellprüfung wurde ein Likelihood-Quotienten-Test (Andersen, 1973, zit. n. Bühner 2011, S. 531). Die χ^2-Prüfgröße ergibt sich daraus, dass der Wert des Likelihood-Quotienten-Tests in eine approximative Prüfgröße überführt wird, welche einer χ^2-Verteilung folgt (vgl. Bühner, 2011, S. 532). Die Stichprobe wird zunächst in zwei Gruppen geteilt. Pro Gruppe werden sodann erneut die spezifischen Item- und Personenparameter geschätzt. Aus den Parameterschätzungen werden

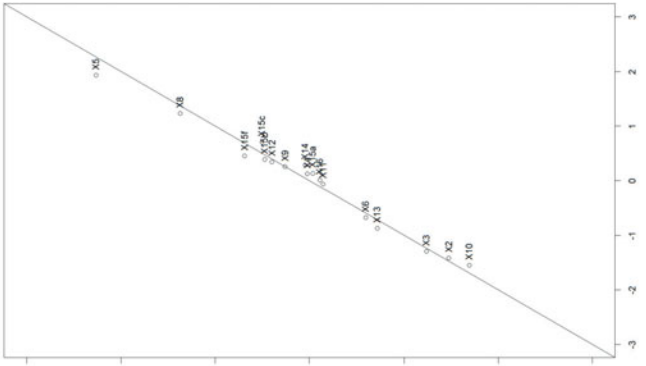
pro Gruppe die Likelihoods bestimmt und der entsprechende Quotient berechnet und in einen empirischen χ^2-Wert überführt. Im Vergleich mit einem kritischen χ^2-Wert, welcher die jeweilige Anzahl an Freiheitsgraden aufweist, kann man testen, ob das Modell abzulehnen ist oder nicht (vgl. ebd.). Das Modell ist zu verwerfen, wenn bei einer Teilung des Datensatzes signifikante Unterschiede zwischen den beiden Teildatensätzen auftreten. Als Teilungskriterium wurden zum ersten ein Median-Split, zum zweiten der arithmetische Mittelwert und zum dritten eine zufällige Teilung herangezogen. Die Ergebnisse des Andersen Tests zeigen, dass das Rasch-Modell im Hinblick auf die ersten beiden Betrachtungen (Mittelwert- und Median-Split) abgelehnt werden muss (siehe Tabelle 16).

Tabelle 16 Ergebnisse des Likelihood-Quotienten-Test

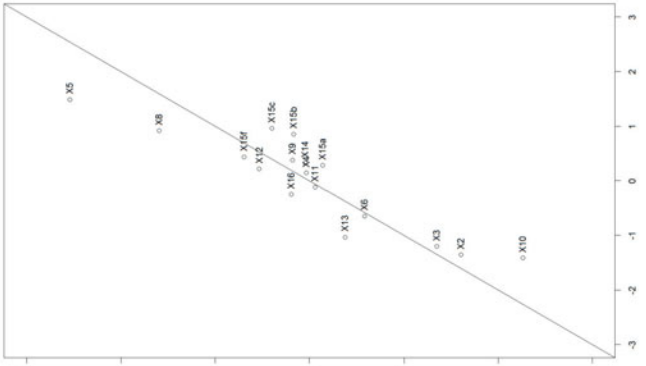
Teilungs-kriterium	Prätest (*N* = 935)			Posttest (*N* = 869)			Follow-up Test (*N* = 922)		
	LR-Wert	*df*	*p*	*LR-Wert*	*df*	*p*	*LR-Wert*	*df*	*p*
Median	50.00	16	< .001	122.19	16	< .001	121.99	16	< .001
Mittelwert	43.43	16	< .001	122.19	16	< .001	121.99	16	< .001
zufällige Teilung	18.44	16	= .299	11.95	16	= .746	9.87	16	= .874

Auch im grafischen Modell-Test zeigen sich zu allen drei Messzeitpunkten in den ersten beiden Betrachtungen (Teilungskriterien: Median und Mittelwert) deutliche Abweichungen, so dass nicht von der Personenhomogenität ausgegangen werden kann (siehe Abbildung 43).

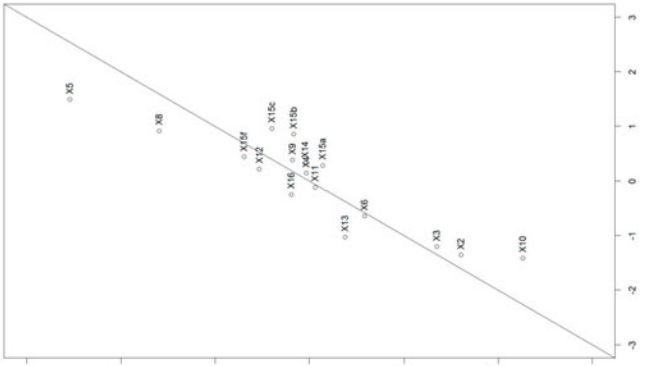
Der grafische Modelltest nutzt die Basisannahme der Personenhomogenität. Sofern das Rasch-Modell zutrifft, misst das zu analysierende Testinstrument in allen beliebigen Teilstichproben dieselbe Fähigkeit. Entsprechend sollten die Schätzungen der Itemparameter für jede Teilstichprobe gleich ausfallen. Grafisch lässt sich diese in einem Streudiagramm darstellen: Fallen die Schätzungen der Itemparameter in den jeweiligen Teilstichproben gleich aus, befinden sich die Itemparameter auf der Diagonalen des Diagramms (vgl. Bühner, 2011, S. 539). Exemplarisch wird hier der grafische Modelltest für den zweiten Messzeitpunkt abgebildet. Für die übrigen beiden Messzeitpunkte ergibt sich jedoch ein ähnliches Bild: Deutliche Abweichungen von der Diagonalen zeigen sich v.a für die Items 5, 10 und 15a-c. Lediglich bei einer zufälligen Teilung kann die Annahme des Raschmodells für den ersten und den zweiten Messzeitpunkt nicht ausgeschlossen werden.



Zufälliges Teilungskriteriium



Beta für die Gruppe, Rohwerte ≥ Mittelwert



Beta für die Gruppe, Rohwerte ≥ Median

Abbildung 43: Grafischer Modelltest: Messzeitpunkt 2 – Posttest

Ein möglicher Grund für die Ablehnung des Modells kann in der Verletzung der Eindimensionalitätsannahme bestehen. Um dies zu prüfen, wurde – entsprechend den Empfehlungen von Bühner (2011) – im nächsten Schritt der Martin-Löf-Test durchgeführt, bei dem es sich um einen „modifizierten Likelihood-Quotienten Test" handelt (vgl. Rost 2004, zit. n. ebd., S. 538). Anstelle die Personen zwei unterschiedlichen Teilstichproben zuzuweisen, werden im Martin-Löf-Test die Items aufgeteilt und anschließend die beiden Testhälften miteinander verglichen. Unter der Voraussetzung, dass das Rasch-Modell gilt, sollten die Items homogen sein und beide Testhälften die gleiche Fähigkeit erfassen. Auch bei diesem Test wird eine empirisch χ^2-verteilte Prüfgröße mit einem kritischen χ^2-Wert verglichen. Fällt die Prüfgröße größer oder gleich als der kritische Wert aus, ist die Eindimensionalitätsannahme verletzt. Als Teilungskriterium wurde für diese Analyse der Median und der Mittelwert der Itemschwierigkeit herangezogen. Wie sich an den Werten ablesen lässt (vgl. Tabelle 17), ist die Annahme der Itemhomogenität zu keinem der drei Messzeitpunkte mit Ausnahme des Konzeptposttests (Teilungskriterium Mittelwert) erfüllt.

Tabelle 17 Ergebnisse des Martin-Löf-Tests

Teilungs-kriterium	Prätest (*N* = 935)			Posttest (*N* = 869)			Follow-upTest (*N* = 922)		
	LR-Wert	*df*	*p*	*LR-Wert*	*df*	*p*	*LR-Wert*	*df*	*p*
Median	57.77	71	= .871	61.09	71	= .793	78.32	71	= .258
Mittelwert	48.10	69	= .974	90.10	69	= .021	68.66	69	= .354

Analog zum Vorgehen beim Leistungstest wurde zudem für jeden der drei Messzeitpunkte getrennt ein Generalfaktorenmodell (konfirmatorische Faktorenanalyse) aufgestellt, bei dem alle Variablen auf einen Faktor laden. Das Generalfaktorenmodell basiert auf einem nachträglich dichotomisierten Datensatz, welcher auch für die Rasch-Analyse verwendet wurde.

Da es sich um ordinalskalierte Daten handelt, wurde der konfirmatorischen Faktorenanalyse eine polychorische Korrelationsmatrix zugrunde gelegt, welche ebenfalls in R berechnet wurde (vgl. Eid et al., S. 922).

Die Ergebnisse der konfirmatorischen Faktorenanalyse bestätigen die Verletzung der Eindimensionalitätsannahme. So musste das Generalfaktorenmodell für jeden der drei Messzeitpunkte verworfen werden. Zum Beleg werden exemplarisch wieder die Fit-Indizes des Generalfaktorenmodells für den zweiten

Messzeitpunkt aufgeführt: $\chi^2_{(119)} = 1388.47$, $p < .001$; $CFI = 0.60$; $TLI = 0.54$; $RMSEA = 0.11$; $SRMR = 0.10$. Die Fit-Indizes der beiden anderen Messzeitpunkte weichen hierbei geringfügig von den Fit-Indizes für den zweiten Messzeitpunkt ab.

Unter Berücksichtigung der vorgestellten Analysen: Likelihood-Quotienten-Test, grafischer Modelltest, Martin-Löf-Test und des Generalfaktorenmodells der konfirmatorischen Faktorenanalyse, muss die Annahme, dass das Raschmodell gilt, insgesamt verworfen werden. Im Hinblick auf eine künftige Weiterentwicklung des Tests werden an dieser Stelle die Ergebnisse für den zweiten Messzeitpunkt (Posttest) berichtet (vgl. Tabelle 18 und Abbildung 44).

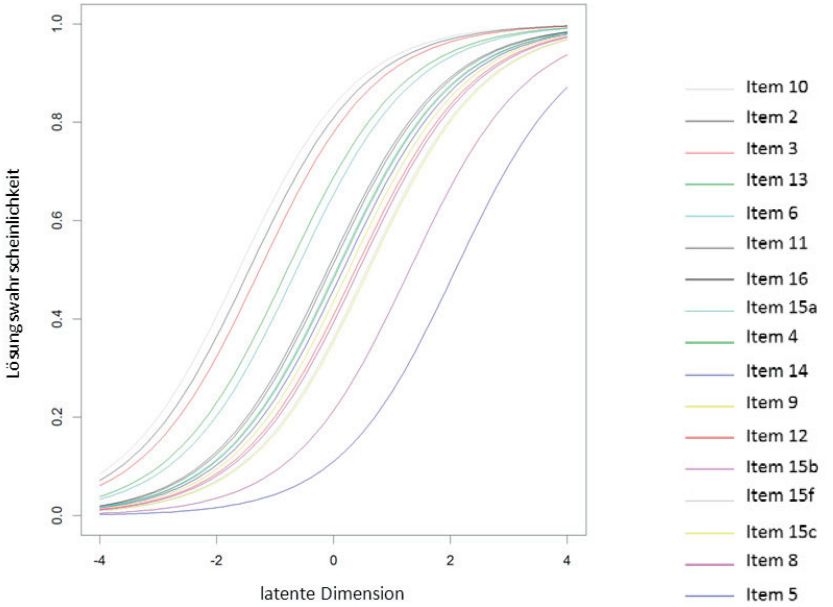


Abbildung 44: Item Characteristic Curve ICC. Die Kurven von links nach rechts entsprechen den Items von oben nach unten (also 10 oben bis 5 unten) in der Legende.

Tabelle 18 Parameterschätzungen der Itemschwierigkeiten für den Konzeptposttest, ($N = 869$)

Item	*Eta*	*SD*	Untergrenze *KI*[a]	Obergrenze *KI*[a]
2	-1.45	-	-	-
3	-1.26	0.07	-1.41	-1.12
4	0.07	0.07	-0.07	0.22
5	2.09	0.13	1.83	2.35
6	-0.63	0.07	-0.77	-0.50
8	1.30	0.10	1.11	1.50
9	0.26	0.08	0.11	0.41
10	-1.62	0.08	-1.77	-1.47
11	-0.10	0.07	-0.25	0.04
12	0.37	0.08	0.22	0.53
13	-0.79	0.07	-0.93	-0.66
14	0.16	0.08	0.01	0.31
15a	0.05	0.07	-0.10	0.20
15b	0.43	0.08	0.28	0.59
15c	0.60	0.08	0.44	0.77
15f	0.57	0.08	0.41	0.73
16	-0.05	0.07	-0.20	0.09

[a]Konfidenzintervall (*KI*) 95 %

Die Normierung des Modells wurde so vorgenommen, dass die Summe der Etas aller Items 0 ergibt (vgl. Tabelle 18). Der Wert des ersten Items (hier Item2), ist somit durch die Werte der anderen Items festgelegt (vgl. Poinstingl et al., 2007).

Die Ergebnisse der vorangegangenen Analysen weisen darauf hin, dass es sich beim Konzepttest um ein mehrdimensionales Testinstrument handelt, das unterschiedliche Fähigkeitsfacetten oder möglicherweise sogar voneinander unabhängige Fähigkeiten erfasst.

Um die Dimensionalität des Tests zu analysieren, wurde das Instrument im nächsten Schritt mittels exploratorischer und konfirmatorischer Faktorenanalyse kreuz validiert.

2.3.5.5 Kreuzvalidierung des Konzepttests

Im Gegensatz zur Rasch-Analyse wurden beiden Analysen der ursprüngliche ordinalskalierte Datensatz zu Grunde gelegt, welcher die Abstufungen 0, 1 und 2 Punkte enthielt, um möglichst viele Informationen beizubehalten.

Das Modell, welches sich aus der exploratorischen Faktorenanalyse ergab, wurde anschließend mittels der Methode der konfirmatorischen Faktorenanalyse überprüft. Um die Ergebnisse der exploratorischen und der konfirmatorischen Faktorenanalyse miteinander in Beziehung setzen zu können, fiel die Wahl in beiden Fällen auf eine Maximum-Likelihood-Fakorenanalyse.

Das Vorgehen umfasste mehrere Schritte:

1. Die Stichproben beider Studien „SV“ und „Ko“ wurden gepoolt ($N_{prä}$ = 935, N_{post} = 869, $N_{follow\text{-}up}$ = 922).
2. Der Datensatz der Gesamtstichprobe wurde zufällig in zwei Hälften geteilt, so dass zu jedem der drei Messzeitpunkte in etwa gleich viele Fälle vorhanden sind. Es wurde geprüft, ob sich die beiden Hälften hinsichtlich der interessierenden Merkmale signifikant unterscheiden. Da dies nicht der Fall war, wurden beide Datensätze für die Kreuzvalidierung verwendet.
3. Es wurde eine exploratorische Faktorenanalyse mit der ersten Hälfte des Datensatzes durchgeführt und zwar zu jedem der drei Messzeitpunkte.
4. Das ermittelte exploratorische Modell wurde an dem zweiten Datensatz je Messzeitpunkt überprüft.
5. Die Schritte 3) und 4) wurden vice versa wiederholt.

Detaillierte Angaben zu den Teilstichproben finden sich in Tabelle 14 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com.

Aus Gründen der Übersichtlichkeit werden an dieser Stelle lediglich die Ergebnisse für den Posttest berichtet. Die Analyseergebnisse konnten für den dritten Messzeitpunkt bestätigt werden (Follow-up Messung). Lediglich zum ersten Messzeitpunkt ergaben sich inhaltliche Abweichungen, welche jedoch angesichts der geringen Vorkenntnisse der Schüler und der daraus resultierenden geringen Reliabilität des Tests nicht überbewertet werden sollten.

Beide Analysen wurden mit der freien Statistiksoftware R unter Verwendung der Pakete psych (Revelle, 2013), lavaan (Rosseel, 2012) und qgraph (Epskamp et al., 2012) durchgeführt. Der Vorteil die freie Statistiksoftware R auch für die Durchführung der exploratorischen Faktorenanalyse zu verwenden, besteht darin, dass R im Gegensatz zu SPSS die Möglichkeit bietet, der Analyse eine polychorische Korrelationsmatrix für ordinalskalierte Daten zugrunde zu legen, welche ebenfalls in R berechnet wurde.

Bei der Durchführung der Analysen zeigte sich, dass (zusätzlich zu den Item 1, 7, 15e-d) weitere Items ausgeschlossen werden mussten, um Doppelladungen zu vermeiden und die Ergebnisse der EFA auch inhaltlich sinnvoll interpretieren zu können. Dies betraf die Items 5, 8, 11, 13 15f, 16 welche bereits in

den Itemstatistiken durch eine vergleichsweise geringe Trennschärfe aufgefallen waren. Von ursprünglich 21 verbleiben nunmehr 11 Items, aus denen sich ein Gesamtscore von maximal 22 Punkten errechnet.

Ausgehend von der Annahme, dass die Faktoren untereinander korrelieren können, wurde in der exploratorischen Faktorenanalyse die Promax-Rotation (Kappa = 4) gewählt. Der Bartlett-Test belegt, dass die Korrelationsmatrix signifikant von der Einheitsmatrix verschieden ist ($\chi^2 = 1072.14$, $df = 55$, $p < .001$) und somit faktorisiert werden kann. Der Verlauf der Eigenwerte und das Ergebnis der Parallelanalyse nach Horn (1965) weisen darauf hin, dass die ersten zwei bis vier Faktoren die Hauptinformationen der Daten repräsentieren (vgl. Abbildung 45).

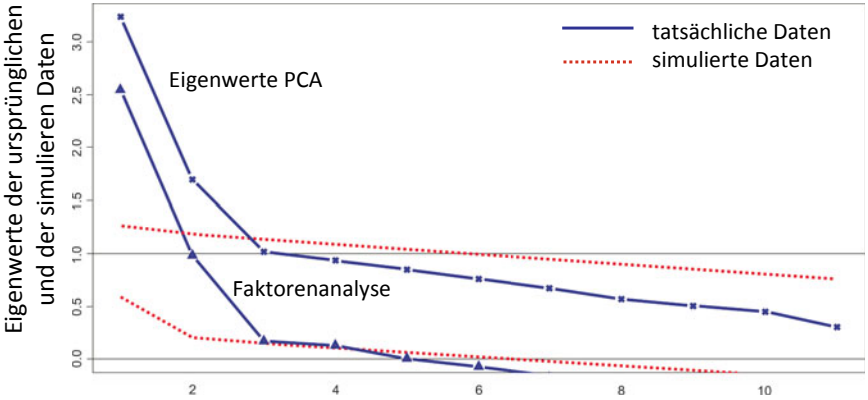


Abbildung 45: Screeplot inklusive Parallelanalyse nach Horn (1965) für den Konzeptposttest

Die Kommunalitäten liegen mit $0.14 \leq 0.75$ in einem für die Stichprobe von $n = 430$ und drei bis fünf Items pro Faktor in einem akzeptablen bis guten Bereich (vgl. Bühner, 2011, S. 345), siehe auch Tabelle 19.

Tabelle 19 Ergebnisse der exploratorischen Faktorenanalyse des Konzeptposttests, ($n = 430$)

Variable		**Promax Rotation Mustermatrix**			**Kommunalität**
y_i	Itemformulierung	F 1	F 2	F3	$\widehat{\boldsymbol{H}}$ (y_i)
y_2	Welche der folgenden Gegenstände / Tiere kann man in einem völlig abgedunkelten Raum sehen?				
	... ein leuchtendes Glühwürmchen	-0.18	**0.42**	0.22	0.22
	... ein weißes Blatt Papier				
	... einen Fahrrad-Reflektor				
	die Augen einer Katze				
y_3	Hat es einen Einfluss auf die Helligkeit in einem Zimmer, ob es helle oder dunkle Tapeten hat?				
	ja, weil die helle Tapete mehr Licht streut, das ins Auge fällt, als eine dunkle Tapete.	0.20	**0.51**	-0.23	0.31
	nein, weil dunkle Tapeten nichts an der Menge des Lichtes im Raum ändern.				
	ja, weil auf der hellen Tapete mehr Licht liegen bleibt.				
	nein, es kommt auf die Lampe in dem Zimmer an oder das Sonnenlicht, das durch das Fenster fällt und nicht auf die Helligkeit der Tapete.				
y_4	Was ist richtig? Lichtstrahlen sind				
	... etwas Wirkliches, so wie dünne Wasserstrahlen aus einer Spritzpistole.	0.10	0.17	**0.20**	0.14
	... etwas Gedachtes, so wie Konstruktionen in der Geometrie, um z.B. Dreiecks-Probleme lösen zu können.				
	... exakt das gleiche wie Lichtbündel.				
	... Lichtbündel sind etwas Gedachtes, z.B. um die Bildgröße bestimmen zu können.				
y_6	Kreuze an, wo sich das Spiegelbild für Dich als Betrachter befindet:				
	... vor dem Spiegel	-0.02	**0.50**	-0.14	0.22
	... hinter dem Spiegel				
	... im Spiegel				
	... auf dem Spiegel				
y_9	In einem abgedunkelten Raum ist der helle Fleck einer Taschenlampe an der Wand zu sehen, nicht aber der Lichtstrahl zwischen Taschenlampe und Wand. Warum?				
	... Erst das an Gegenständen gestreute Licht trifft ins Auge und ist sichtbar.	0.01	**0.45**	-0.01	0.20
	... In dem dunklen Raum wird das Licht absorbiert (verschluckt), daher ist es nicht zu sehen.				
	... Das Licht erhellt die Wand, weil es auf ihr liegen bleibt.				
	... Das Licht der Taschenlampe entfernt sich vom Beobachter, erst durch dieWand wird es umgedreht und geht auf den Beobachter zu.				
y_{10}	Was passiert, wenn man in dem Lichtstrahl einen Tafellappen aufschüttelt?				
	... Die Staubteilchen wirken wie kleine Linsen, die das Licht auf der Wand bündeln.	-0.06	**0.67**	0.16	0.52
	... Der feine Kreidestaub sammelt das Licht und dadurch sieht man den hellen Fleck auf der Wand nicht mehr.				
	... Die Staubteilchen werden durch das auftreffende Licht durcheinander gewirbelt.				

	... Die Staubteilchen streuen das Licht in alle Richtungen, dadurch trifft es ins Auge und wird sichtbar.				
y_{11}	Wie entsteht durch Verwendung einer Sammellinse ein Bild, das auf einem Schirm aufgefangen werden kann?				
	... Solch ein Bild entsteht durch Spiegelung der Lichtstrahlen an der Linse nach dem Reflexionsgesetz.				
	... Eine Sammellinse hat den Effekt, die Lichtstrahlen aufzuhellen.	0.06	0.10	**0.29**	0.14
	... Lichtstrahlen, die von einem Gegenstandspunkt ausgehen, werden durch die Sammellinse abgelenkt und treffen sich im Bildpunkt.				
	... Das Bild geht als Ganzes durch die Linse zum Schirm, dabei wird es in der Linse unter Einhaltung der Linsengesetze umgedreht.				
y_{14}	Welche Aussagen zur Bildkonstruktion und Bildentstehung treffen zu?				
	... Nur die ausgezeichneten Strahlen kann man im Strahlengang zeichnen.				
	... Mit den ausgezeichneten Strahlen kann man den Strahlengang besonders leicht zeichnen.	0.05	-0.03	**0.69**	0.50
	... Die ausgezeichneten Strahlen erschweren die Zeichnung, machen sie dafür aber besonders genau.				
	... Ohne die ausgezeichneten Strahlen (wenn diese z.B. durch dünne Stifte aufgehalten werden) kann es kein Bild geben.				
y_{15a}	In einer Versuchsanordnung sind eine Glühlampe, eine Sammellinse und ein Schirm auf einer optischen Bank so montiert, dass ein vergrößertes, umgekehrtes, scharfes Bild des Glühfadens entsteht: Was passiert, wenn die untere Hälfte der Linse abgedeckt.				
	... Die obere Hälfte des Bildes wird abgeschnitten.	**0.72**	-0.08	0.13	0.58
	... Die untere Hälfte des Bildes wird abgeschnitten.				
	... Das Bild wird dunkler.				
	... Das Bild wird kleiner.				
y_{15b}	Was passiert, wenn man einen Karton mit großem Loch (ringförmige Blende) vor die Linse hält?				
	... Das Bild wird kleiner.				
	... Das Bild wird dunkler.	**0.88**	0.01	-0.03	0.75
	... Die Ränder des Bildes werden kreisförmig abgeschnitten.				
	... Das Bild wird heller.				
y_{15c}	Was passiert, wenn man einen Karton mit einem sehr kleinen Loch 5 mm (ringförmige Blende) vor die Linse hält?				
	... Das Bild wird kleiner.				
	... Das Bild wird dunkler.	**0.73**	0.09	0.05	0.64
	... Die Ränder des Bildes werden kreisförmig abgeschnitten.				
	... Das Bild wird heller.				

TLI = 0.91, *RMSR* = 0.03, *RMSEA* = 0.06

Es zeigt sich, dass die Items 15a-c einen höheren und die Items 4, 11 und 14 einen eher geringeren Varianzanteil aufklären (vgl. Tabelle 20).

Tabelle 20 Eigenwerte der Faktoren, Reliabilität und Anteil aufgeklärter Varianz der resultierenden Skalen für die Faktorenanalyse des Konzeptposttests, (n = 430)

F[a]	Interpretation	Variablen	α	Anfänglicher Eigenwert	Anteil erklärter Varianz in %	Anteil kumulierter Varianz in %
1	Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben	y_{15a}, y_{15b}, y_{15c}	.81	3.39	18	18
2	Geradlinige Lichtausbreitung und Streuung	y_2, y_3, y_6, y_9, y_{10}	.65	1.78	13	31
3	Verständnis Bildkonstruktion / Strahlenmodell	y_4, y_{11}, y_{14}	.49	1.37	7	38

[a]Faktor

Insgesamt werden 38 % der Varianz durch die drei Faktoren aufgeklärt.

Im zweiten Schritt wurden die Ergebnisse der exploratorischen Faktorenanalyse nun mit der zweiten Hälfte des Datensatzes (n_{post} = 430) kreuzvalidiert.

Auf Basis des Modells, das sich an der exploratorischen Faktorenanalyse orientierte, waren die Items, wie folgt, zugeordnet:

- Faktor 1 („Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben“) = y_{15a}, y_{15b}, y_{15c}
- Faktor 2 („geradlinige Lichtausbreitung und Streuung“) = y_2, y_3, y_6, y_9, y_{10}
- Faktor 3 („Verständnis Bildkonstruktion / Strahlenmodell“) = y_4, y_{11}, y_{14}

Um zu analysieren, wie die verschiedenen Skalen zusammenspielen, wurde das Modell nun auf unterschiedliche Weise spezifiziert:

- Basismodell mit Korrelationen zwischen Faktoren: Zunächst wurde ein Modell aufgestellt, bei dem die jeweiligen Items auf einen Faktor laden. Das Programm schätzte dabei Korrelationen zwischen den drei Faktoren.
- Bifaktormodell: In einem nächsten Schritt wurde ein bifaktorielles Modell aufgestellt. In diesem Modell wird angenommen, dass ein Generalfaktor existiert, der die Bearbeitung aller Items beeinflusst. Darüber hinaus bestehen jedoch auch latente Merkmale, die über das Merkmal, welches eine übergreifende Kompetenz misst, hinaus auch noch voneinander unabhängige Merkmale erfassen. Korrelationen zwischen den Faktoren wurden daher unterbunden.
- Second-Order-Modell: Im letzten Schritt wurde ein Second-Order-Modell aufgestellt. Dieses Modell geht davon aus, dass ein „generelles“ latentes Merkmal besteht, welches sich wiederum auf alle spezifischen latenten Merkmale auswirkt, die voneinander unabhängig und daher untereinander unkorreliert sind.

Tabelle 21 Konfirmatorische Faktorenanalyse Konzeptposttests – Modellvergleich des empirisch aus der Faktorenanalyse abgeleiteten Modells (n = 439)

Kriterium (globale Fitindizes)	Basismodell mit Korrelationen	Bifaktorielles Modell	Second-Order-Modell
$\chi^2_{(df)}$	104.17 $_{(41)}$*	104.27 $_{(41)}$*	104.34 $_{(42)}$*
CFI	0.94	0.94	0.94
TLI	0.92	0.92	0.92
RMSEA	0.06	0.08	0.06
SRMR	0.05	0.07	0.05
AIC	12761.63	12761.63	12759.70
BIC	12863.74	12863.74	12857.72

*p < .05

Ein Vergleich der Fit-Indizes zeigt, dass alle Modelle sehr ähnliche globale Fit-Indizes aufweisen (vgl. Tabelle 21). Die angegeben globalen Fit Indizes der Modelle liegen jeweils unter bzw. über den in der Literatur angegebenen Cut-off-Werten (vgl. Bühner, 2011, S. 428). Nach dem Akaike- und dem Bayes-Informationskriterium schneidet jeweils das Second-Order-Modell am besten ab (geringste Werte). Hinsichtlich der lokalen Fit-Werte erwies sich die Modellspezifikation, bei der die jeweiligen Items auf einen Faktor laden und die Faktoren untereinander korrelieren, am geeignetsten (Basismodell mit Korrelationen), da im Gegensatz zu der bifaktoriellen Variante und dem Second-Order-Modell jeder der betreffenden Pfadkoeffizienten signifikant ist. Die Varianzaufklärung je Item unterscheidet sich zwischen Second-Order-Modell und Basismodell erst ab der dritten Nachkommastelle. Die lokalen Fit-Indizes sprechen insgesamt für das Basismodell mit Korrelationen.

Letzteres kann inhaltlich wie folgt interpretiert werden (vgl. Abbildung 46): Es bestehen drei Dimensionen des konzeptuellen Verständnisses im Bereich der Strahlenoptik:

- AB: Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben,
- LS: geradlinige Lichtausbreitung und Streuung,
- BS: Verständnis Bildkonstruktion / Strahlenmodell.

Alle drei Dimensionen tragen zum Verständnis der Bildentstehung mittels Sammellinse bei. Sie hängen inhaltlich miteinander zusammen, was sich in den wechselseitigen signifikanten Korrelationen zeigt. Dabei liegen die Korrelationen ($0.42 \leq r \leq 0.71$) im mittleren bis hohen Bereich (siehe ebenfalls Abbildung 46).

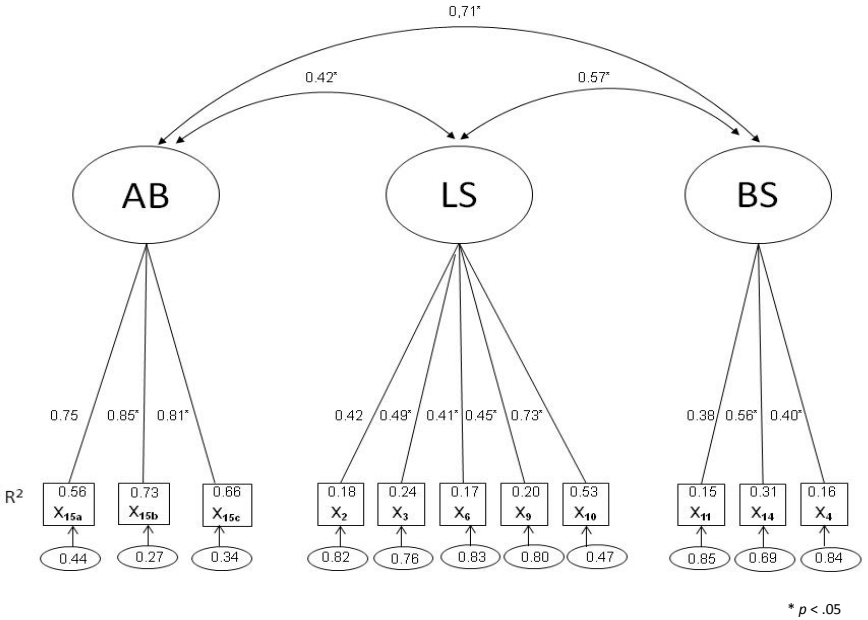


Abbildung 46: Konfirmatorische Faktorenanalyse aus der EFA empirisch abgeleitetes Basismodell mit Korrelationen: standardisierte Lösung, AB = Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben, LS = geradlinige Lichtausbreitung und Streuung, BS: Verständnis Bildkonstruktion / Strahlenmodell,

Während die globalen Gütekriterien in einem guten bis akzeptablen Bereich liegen, sind die lokalen Gütekriterien der Indikatorreliabilität und der diskriminanten Validität (Formel-Larker-Kriterium) für die beiden Faktoren „LS: geradlinige Lichtausbreitung und Streuung“, und „BS: Verständnis Bildkonstruktion / Strahlenmodell“ nicht zufriedenstellend erfüllt (vgl. Tabelle 22).

Als Anhaltspunkt für die Kriteriumsvalidität wurden des Weiteren Korrelationen zwischen der Gesamtpunktzahl aus den verbleibenden 11 Items und den Fachnoten sowie relevanten Subskalen des I-S-T 2000 R für die letzten beiden Messzeitpunkte berechnet.

Im Ergebnis zeigen sich signifikante mittlere Korrelationen zwischen den Fachnoten in Physik (r_{post_PH} (479) = .42, $p < .001$, $r_{follow\text{-}up\text{-}PH}$ (476) = .38, $p < .001$) und Mathematik (r_{post_Mathe} (476) = .38, $p < .001$, $r_{follow\text{-}up\text{-}Mathe}$ (483) = .40, $p < .001$) und den Gesamtpunktzahlen im Post- und Follow-up Test. Es ergeben sich jedoch nur

geringe Korrelationen ($r < .15$) zwischen der Gesamtpunktzahl post und follow-up und den Subskalen des figural-räumlichen und figural-logischen Schlussfolgern.

Tabelle 22 Lokale Gütekriterien der CFA des Konzeptposttests, ($n = 439$)

F	Interpretation	Item	Indikatorreliabilität	Faktor-reliabilität	*DEV*	Formel Larcker Kriterium
AB[a]	Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben	y_{15a}	0.57	.83	0.62	> 0.50
		y_{15b}	0.73			
		y_{15c},	0.66			
LS[b]	Geradlinige Lichtausbreitung und Streuung	y_2	0.18	.77	0.42	.64 (nicht erfüllt)
		y_3	0.24			
		y_6	0.17			
		y_9	0.20			
		y_{10}	0.53			
BS[c]	Verständnis Bildkonstruktion / Strahlenmodell	y_4	0.15	.55	0.30	.50 (nicht erfüllt)
		y_{11}	0.31			
		y_{14}	0.16			

[a]AB = Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben
[b]LS = geradlinige Lichtausbreitung und Streuung
[c]BS = Verständnis Bildkonstruktion / Strahlenmodell

2.3.5.6 Itemstatistiken zum Motivationsfragebogen

Der Motivationsfragebogen basiert auf Items, die sich in bisherigen Studien bewährt haben. Somit handelt es sich anders als beim Leistungstest und beim Konzepttest nicht um ein im Rahmen des Projekts entwickeltes Instrument, sondern um eine auf die Fragestellung zugeschnittene Auswahl und Zusammenstellung von Skalen.

Um das Zusammenspiel dieser Skalen zu untersuchen, wird die Itemanalyse des Motivationsfragebogens im Überblick vorgestellt, zumal die Analyse der Wirkung der Intervention auf die Motivation einen Nebenaspekt dieser Arbeit darstellt. Zur Berücksichtigung der Schülermotivation wurden in die Gesamtpunktzahl die zuvor genannten Skalen SK (Selbstkonzept), IE (Intrinsische Motivation / Engagement und GN (Gute Noten - extrinsische Motivation) einbezogen. Die Skala LES: „wahrgenommenes Lehrerengagement aus Schülersicht“ (abgewandelt

nach Seidel et al., 2003) wurde aus der Gesamtwertung der Schülermotivation herausgenommen und ging als gesonderte Kovariate in die Analyse ein.

Die Analyse der Itemschwierigkeit und Mittelwerte zeigt, dass alle Items (inklusive der Items zu LES) zu allen drei Messzeitpunkten in einem guten Bereich von $0.20 \leq P_i \leq 0.80$ bezüglich der Itemschwierigkeit lagen (vgl. Tabelle 15 und 16 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com). In Bezug auf die korrigierte Trennschärfe ergaben sich zufriedenstellende bis gute Werte. So liegt das Minimum bei $r_{it} = .28$ im Follow-up Test und das Maximum jeweils bei $r_{it} = .79$ im Posttest (vgl. Tabelle 16 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com). Cronbachs Alpha als Maß für die interne Konsistenz und Schätzer für die Reliabilität kann für die Schülermotivation und LES als hoch bewertet werden. Für die Test-Retest-Reliabilität ergeben sich für beide Konstrukte zufriedenstellende bis gute Werte (vgl. Tabelle 23).

Tabelle 23 Reliabilitätsstatistiken für den Motivationsfragebogen

Reliabilitätschätzung	Schülermotivation (24 Items)			Lehrerengagement aus Schülersicht (4 Items)		
	prä (*N* = 387)	post (*N* = 362)	follow-up (*N* = 391)	prä (*N* = 465)	post (*N* = 451)	follow-up (*N* = 471)
α	.93	.94	.94	.88	.90	.91
$r_{Test\text{-}Retest}$	prä – post	prä – follow-up	post – follow-up	prä – post	prä – follow-up	post – follow-up
	.76***	.69***	.77***	.67***	.63***	.67***

***$p < .001$

Abweichende Werte für die Stichprobengrößen zwischen Schülermotivation und Lehrerengagement ergeben sich aus dem listenweisen Fallausschluss, bei dem komplette Bögen ausgeschlossen wurden, sobald mindestens ein Item unbeantwortet blieb. Für die späteren Analysen wurde in diesen Fällen Daten imputiert (siehe Kapitel 2.3.9.1 Methodisches Vorgehen: Mehrebenenanalyse).

2.3.5.7 Diskussion der Ergebnisse

Diskussion der Analysen zum Leistungstest: Die Sicherung der Gütekriterien Objektivität, Reliabilität und Validität des Leistungstests zur Erfassung der Physikleistung bei repräsentationsbezogenen Aufgaben kann zusammenfassend als zufriedenstellend bis gut bewertet werden.

Die Sicherung der Durchführungsobjektivität ergibt sich aus dem standardisierten Verfahren der Instruktion der Schüler. So wurde der Test jeweils von der Autorin durchgeführt und die Durchführung in gleicher Weise in den verschiedenen Schulklassen erklärt. Die Auswertungsobjektivität wurde durch ein zuvor festgelegtes Bewertungsschema gewährleistet, in dem detailliert festgelegt wurde, welche Teilpunkte für welche Antworten oder Bestandteile der Strahlenkonstruktion vergeben waren, an welche sich die beteiligten Auswerter einschließlich der Autorin hielten.

Die Analyse der Itemstatistiken des Leistungstests belegen zufriedenstellende bis gute Werte der Kennwerte (Lösungswahrscheinlichkeit, Trennschärfe und Cronbachs-Alpha, Test-Retest-Reliabilität) für den zweiten und den dritten Messzeitpunkt. Daher kann die Reliabilität als insgesamt zufriedenstellend bis gut eingestuft werden.

Da die Schüler zuvor keinen Unterricht zur Bildentstehung bei der Sammellinse erhalten hatten, waren die niedrigen Werte in den Itemstatistiken (Mittelwert, Standardabweichung, Lösungswahrscheinlichkeit und Trennschärfe) zum ersten Messzeitpunkt zu erwarten. Für die Beurteilung, welche Items beibehalten werden, wurden daher vorwiegend die Daten der Post- und der Follow-up Messung in Betracht gezogen.

Die Items 2b und 6 zeichneten sich durch eine geringere Lösungswahrscheinlichkeit aus. Inhaltlich beziehen sich die beiden Items darauf, aus der schematischen Repräsentation der Strahlenkonstruktion Schlussfolgerungen zu ziehen und diese in Worte zu fassen.[30] Spezielle Voraussetzung für die Bearbeitung der Aufgaben sind Kenntnisse in Geometrie. Offenbar wurden die hierzu benötigten Fähigkeiten im Unterricht nicht im ausreichenden Maß vermittelt und eingeübt. Die Lösungswahrscheinlichkeit für Item 2b liegt hierbei mit 18 % in der Postmessung knapp unter der kritischen Grenze. Ergänzend muss jedoch hinzugefügt werden, dass es sich in beiden Fällen um Transferaufgaben handelt, die ein hohes Maß an Repräsentationskompetenz erfordern und durch unterschiedliche Argumentationswege bearbeitet werden können. Vor diesem Hintergrund war eine geringe Lösungswahrscheinlichkeit erwartbar. So ist aus der Forschung zur Problemlösekompetenz bekannt, dass Schülern die Bearbeitung von Transferaufgaben schwerfällt (vgl. Funke & Zumbach, 2006; S. 216 f. im Überblick).

30 Wortlaut des Items 2b: „Begründe mit den Strahlen in der Abbildung oben: Warum verändert sich die Bildgröße so wie oben beschrieben, wenn der Gegenstand weiter von der Linse entfernt wird?“; Wortlaut des Items 6 „Warum kann man die fett hervorgehobenen Dreiecke nicht zur Herleitung der Abbildungsgleichung verwenden?“

Als weiterer bemerkenswerter Befund zeigte sich, dass die Lösungswahrscheinlichkeit für die Wahl der korrekten Gegenstandsweite (5c - Erklärung1) deutlich weniger stark sinkt als die Lösungswahrscheinlichkeit für die Erklärung dieser Wahl (5c - Erklärung2). Gründe könnten zum einen darin liegen, dass sich die Schüler daran erinnern, an welche Stelle sie das Kreuz in der Multiple-Choice Aufgabe gesetzt haben, aber die Begründung nicht mehr herleiten können, da die Inhalte nicht nachhaltig erlernt wurden, oder aber, dass zum dritten Messzeitpunkt die Motivation fehlte, wiederholt eine konsistente Begründung zu entwickeln, zumal der Follow-up Test nicht in die Notengebung einfließt. Gegen die zweite Hypothese spricht, dass die Begründung für Aufgabenteil 5a häufiger korrekt entwickelt wurde. Da die Itemschwierigkeit für Item 5c - Erklärung 2 in der Postmessung jedoch innerhalb des Toleranzbereichs liegt, wurde das Item vorerst beibehalten.

Bei der Analyse der Trennschärfe fielen die Ergebnisse für Item 7 (Anwendung der Abbildungsgleichung) auf. Schüler, die im Vortest Item 7 zumindest teilweise richtig lösten, bearbeiteten die übrigen Aufgaben im Vortest offenbar mit weniger Erfolg. Im Posttest steigt die Trennschärfe für dieses Item deutlich an und sinkt im Follow-up Test wieder. Eine mögliche Erklärung könnte darin bestehen, dass die Bearbeitung dieser klassischen Aufgaben in hohem Maß von erlerntem Wissen beeinflusst wird. Schüler, die im Vortest möglicherweise „raten“, erzielten hierbei ggf. eine Teilpunktzahl durch die Angabe der Gleichung, die ihnen z.B. durch die Lochkamera bekannt sein könnte. Andere Schüler hingegen, die erlerntes Wissen anwenden, ließen die Beantwortung des Items aus, so dass sich Item 7 bei Schülern mit sehr geringen Vorkenntnissen nicht dazu eignet, zwischen leistungsstarken und leistungsschwachen Schülern zu differenzieren. Da zwischen der Post- und der Follow-up Messung ein Zeitabstand von ca. zwei Monaten gegeben war, wurde vermutlich der Lerninhalt der Abbildungsgleichung vergessen, obgleich gegebenenfalls das Grundverständnis für die Bildentstehung erhalten blieb, was den Rückgang der Trennschärfe im Follow-up Test erklären könnte.

Die Analyse der Itemstatistiken und die Ergebnisse der Faktorenanalyse legten nahe, Aufgabe 6 („Warum kann man die fett hervorgehobenen Dreiecke nicht zur Herleitung der Abbildungsgleichung verwenden?“) von der Gesamtwertung auszuschließen, auch wenn sich hierdurch Cronbachs Alpha nicht veränderte.

Die Ergebnisse der Faktorenanalyse weisen darauf hin, dass der Test insgesamt drei Kernkompetenzen erfasst, die sich auf den Umgang mit verschiedenen Repräsentationsformen beziehen: den Umgang mit deskriptiven Repräsentationen, die Aufgabe, auf Basis einer realistisch depiktionalen Repräsentation Bezüge zu einem realen Experiment herzustellen und auf den Umgang mit schematisch depiktionalen Repräsentationen. Die Modellvergleiche belegen, dass die Art der

Repräsentationsform entscheidender ist als die Unterscheidung, ob Schülervorstellung thematisiert werden sollen oder nicht (Vergleich der 2-Faktoren- mit den 3-Faktorenmodellen).

Die guten globalen Fitwerte des Bifaktormodells und die hohen Korrelationen der Faktoren im Basismodell weisen darauf hin, dass die drei Kernkompetenzen vermutlich durch eine generelle „Repräsentations"-Kompetenz oder ein physikalisches Grundverständnis der Bildentstehung beeinflusst werden.

Die Struktur des Tests ist mit den Ausführungen im Theorieteil dieser Arbeit, insbesondere mit den Ausführungen zur Funktion multipler Repräsentationen (siehe Kapitel 1.1.4.1 Funktionen von multiplen Repräsentationen) konform, was als Beleg für die Konstruktvalidität gewertet werden kann. Die Überlegenheit der 3-Faktorenmodelle gegenüber den 2-Faktorenmodellen stützt die Bedeutung der Unterscheidung der repräsentationalen Anforderungen.

Zur Optimierung der lokalen Fit-Werte sollten insbesondere die Aufgaben zur Erfassung des Umgangs mit schematischen Repräsentationen überarbeitet werden, um die interne Konsistenz dieser Skala zu erhöhen. Zudem könnte die Skala um ein bis zwei Aufgaben erweitert werden, welche einen Transfer bei der Bearbeitung der Strahlenkonstruktion erfordern. Weitere Aufgaben, welche in dieser Studie nicht zum Einsatz kamen, finden sich in Scheid (2013).

Die curriculare Validität des Tests wurde in einem Expertenrating von Scheid (2013, S. 139 - 140) untersucht. An dem Rating nahmen 11 Gymnasiallehrer mit einer durchschnittlichen Unterrichtserfahrung von $M = 20.91$ Jahren ($SD = 13.46$ Jahre) teil, deren Urteile nach Kendalls Konkordanzkoeffizient W höchst signifikant übereinstimmten (vgl. Wirtz & Caspar, S. 155). Im Ergebnis zeigte sich, dass die Items bis auf zwei Ausnahmen, als curricular valide eingestuft wurden. Lediglich die Aufgaben 5c) und 6), wurden als für einen Leistungstest weniger geeignet bzw. als weniger curricular valide eingeschätzt und wären entsprechend von den Lehrern eher nicht im eigenen Unterricht verwendet worden (vgl. Scheid, 2013, S. 139 - 141).

2.3.5.8 Fazit zur ersten Zielsetzung: Entwicklung eines tragfähigen Konzepttests

Die Sicherung der Gütekriterien Objektivität, Reliabilität und Validität des Konzepttests zur Erfassung des konzeptuellen Verständnisses in der Strahlenoptik kann zusammenfassend ebenfalls als zufriedenstellend bis gut beurteilt werden.

Die Sicherung der Durchführungsobjektivität ergibt sich analog zum Vorgehen beim Leistungstest aus dem standardisierten Verfahren der Instruktion, welches

jeweils von der Autorin durchgeführt wurde. Die Auswertungsobjektivität wurde durch ein zuvor festgelegtes Bewertungsschema sichergestellt (siehe Kapitel 2.3.4.3 Konzepttest).

Für die Analyse der Itemstatistiken und der internen Konsistenz ergeben sich unter Ausschluss der genannten Items zufriedenstellende bis gute Werte für den zweiten und den dritten Messzeitpunkt (Lösungswahrscheinlichkeit, Trennschärfe und Cronbachs-Alpha, Test-Retest-Reliabilität) – zumal zu berücksichtigen ist, dass das konzeptuelle Verständnis von Schülern, die keine Experten sind, auch nach dem Unterricht nach wie vor von Inkonsistenzen geprägt sein dürfte. Die niedrigeren Werte im Prätest resultieren aus dem geringen Vorwissen der Schüler.

In der Analyse der Struktur belegen die Untersuchungen zur Prüfung auf Rasch-Skalierbarkeit für den nachträglich dichotomisierten Datensatz und die Prüfung der Eindimensionalitätsannahme für den ordinalskalierten Datensatz, dass es sich beim konzeptuellen Verständnis um ein mehrdimensionales Konstrukt handelt. Entsprechend kann der Test nicht Rasch-skaliert werden.

Die Idee der Rasch-Skalierung kann jedoch für die künftige Weiterentwicklung des Tests aufgegriffen werden, unter der Voraussetzung, dass insbesondere die Items 2, 3, 5, 8 und 10, welche sich vorwiegend auf Konzepte zur Lichtausbreitung und Streuung beziehen, ausgeschlossen oder deutlich überarbeitet werden.

In der Kreuzvalidierung (EFA und CFA) zeigte sich, dass der Test insgesamt drei übergeordnete Verständnisbereiche erfasst:

- Bildentstehung inklusive Abdeckaufgaben,
- geradlinige Lichtausbreitung und Streuung,
- Bildkonstruktion und Strahlenmodell.

Die Modellvergleiche belegen, dass die drei Verständnisbereiche unterschiedliche Facetten des konzeptuellen Verständnisses erfassen, welche jedoch miteinander korrelieren. Die Korrelationen sprechen für einen mittleren bis hohen Zusammenhang der drei Facetten untereinander. Die höchste Korrelation besteht zwischen den beiden Faktoren „Abdeckaufgaben“ und „Bildkonstruktion Strahlenmodell“, was als empirischer Anhaltspunkt für Wiesners Annahme (Wiesner, 1992a, 1992b) interpretiert werden kann, dass Schüler, welche Lernschwierigkeiten bei Abdeckaufgaben haben, das Konzept der Punkt-zu-Punkt-Abbildung nicht verstanden haben.

Zur Optimierung der lokalen Fit-Werte in der CFA, sollten insbesondere die Aufgaben zur Erfassung der Bildkonstruktion und des Strahlenmodells überarbeitet werden. Diese Skala könnte um ein bis zwei weitere Aufgaben erweitert werden,

welche speziell das Konzept der Punkt-zu-Punkt-Abbildung thematisieren. Die in sich schlüssigen Ergebnisse und die akzeptablen globalen Fitwerte des Modells sprechen dafür, dass das Gütekriterium der Konstruktvalidität (interne Validität) ausreichend erfüllt ist.

Die signifikanten mittleren Korrelationen mit den Fachnoten in Physik und Mathematik belegen den Zusammenhang der erfassten Konstrukte mit dem relevanten Außenkriterium der Schulleistung in inhaltlich nahen Fächern (Kriteriumsvalidität).

Als Resümee kann festgehalten werden, dass die verbleibenden 11 Items eine tragfähige Testfassung bilden. Neben der Erweiterung der Skala „Bildkonstruktion und Strahlenmodell" zur Optimierung der internen Konsistenz und internen Konstruktvalidität, könnte die (externe) Konstruktvalidität durch weitere Kriterien – wie offene Interviews mit Teilnehmern, welche Hinweise auf die angewandten Lösungsstrategien bieten – und Expertenratings zur Beurteilung der inhaltlichen Validität des Tests abgesichert werden (vgl. Lindell, Peak & Foster, 2007).

Fazit zu den Analysen des Fragebogens zu Erhebung der Motivation und des Lehrer-Engagments aus Schülersicht: Für die Erhebung der Schülermotivation und des LES ergeben sich gute bis sehr gute Schätzwerte für die Reliabilität. Zur Diskussion der Validität sei auf die genannte Originalliteratur verwiesen aus der die Items stammen (Rheinberg & Wendland 2003, 2004; Seidel et al., 2001).

2.3.6 Ergebnisse zu den Kovariaten

2.3.6.1 Ergebnisse zu kognitiven Fähigkeiten

Zur Erfassung der kognitiven Fähigkeiten wurden drei Einzelwerte erhoben: verbales Schlussfolgern mit der Aufgabengruppe Satzergänzung (SE), figural-räumliches Schlussfolgern mit der Aufgabengruppe Würfelaufgaben (WÜ) und figural-logisches Schlussfolgern mit der Aufgabengruppe Matrizen (MA). Da es sich bei dem Intelligenztest um ein normiertes Instrument handelt, wurde auf detaillierte Angaben zur Itemanalyse verzichtet und pro Skala der Range angegeben. Die Analyse der Itemschwierigkeit zeigt, dass die Items aller drei Skalen in etwa in dem Bereich liegen, welchen die Autoren angeben. Gleiches gilt für die Trennschärfe. Bezüglich beider Kennwerte zeigte sich aber auch, dass die Werte für die Trennschärfe und die Lösungswahrscheinlichkeit v.a. für die jeweils letzte und schwerste Aufgabe der Skala abweichen, da diese Aufgaben nur von wenigen Schülern gelöst wurden (vgl. Tabelle 24).

Tabelle 24 Vergleich der Schätzungen der Reliabilität: gemessene Werte und Werte von Liepmann et al. (2007) im I-S-T 2000 R

Aufgabengruppe	Itemstatistiken	Werte „SV“ (N = 479)[a]	Werte der Autoren (n = 1445)[b]
Satzergänzung (20 Items)	Lösungswahrscheinlichkeiten (Range)	.13 - .82	.19 - .94
	Trennschärfen (Range)	.00 - .43	.09 - .43
	Cronbachs Alpha	.60	.63
	Split Half	.48	.67
Würfelaufgaben (20 Items)	Lösungswahrscheinlichkeiten (Range)	.08 - .88	.10 - .82
	Trennschärfen (Range)	.00 - .49	.23 - .52
	Cronbachs Alpha	.74	.79
	Split Half	.50	.84
Matrizen (20 Items)	Lösungswahrscheinlichkeiten (Range)	.15 - .93	.06 - .96
	Trennschärfen (Range)	.05 - .34	.15 - .41
	Cronbachs Alpha	.54	.66
	Split Half	.44	.70

[a] Gesamtstichprobe „SV“: Schülervorstellungen (N = 479, n_{Gy} = 399, n_{IGS} = 80)
[b] Vergleichswerte für die Teilstichprobe gymnasial (vgl. Liepmann et al, 2007, S. 24 ff.)

Tabelle 25 Mittelwert und Standardabweichung der Subskalen des I-S-T 2000 R nach Bedingung und Schultyp, (N = 479)

Skala	TG[a]		KG[b]		Gymnasium		IGS[c]	
	(n = 220)		(n = 229)		(n = 399)		(n = 80)	
	M	*SD*	*M*	*SD*	M	SD	M	SD
SE[d]	99.34	10.65	98.93	10.55	100.80	10.32	90.90	7.70
WÜ[e]	97.70	9.14	98.92	9.06	99.09	8.98	94.28	8.74
MA[f]	100.52	9.44	101.19	9.66	101.74	9.62	96.33	7.77

[a]Treatmentgruppe, [b]Kontrollgruppe, [c]Integrierte Gesamtschule, [d]Satzergänzung, [e]Würfelaufgaben, [f]Matrizen

Cronbachs Alpha als Maß für die interne Konsistenz und Schätzer für die Reliabilität kann als ausreichend bewertet werden, liegt jedoch für die drei verwendeten Skalen unter den Werten, welche die Autoren des Tests angeben (vgl. ebenfalls Tabelle 24).

Tabelle 17 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com enthält eine ausführliche Aufstellung deskriptiver Statistiken bezüglich der drei Subskalen. Die Mittelwerte der drei Intelligenzkomponenten der Gymnasiasten hinweg liegen nahe bei 100 und die Standardabweichungen nahe bei 9 (vgl. Tabelle 25). Dies entspricht den Werten der Autoren, welche den Test

auf einen Mittelwert von $M = 100$ und eine Standardabweichung von $SD = 10$ normiert haben (vgl. ebd., S. 40). Die im Rahmen der Studie untersuchte Gruppe entspricht in etwa knapp dem Bevölkerungsschnitt. Abweichungen in den Mittelwerten der Skalen nach unten ergeben sich vor allem für die Gesamtschüler.

Zur grafischen Veranschaulichung der Daten wurden Boxplots der drei Intelligenzkomponenten nach Bedingung (TG versus KG) und Schultyp (Gymnasium versus IGS) erstellt (vgl. Abbildungen 47 und 48). Auf deskriptiver Ebene zeigen sich in den Boxplots im verbalen Bereich höhere Werte in der Treatmentgruppe: So ist der Median nach oben verschoben. Im Bereich des figural-räumlichen Denkens verhält es sich genau umgekehrt. Hier schnitten die Schüler der Kontrollgruppe etwas besser ab. Zudem ist die Streuung der Werte in diesem Bereich in der Treatmentgruppe höher (vgl. Abbildung 47). Im Bereich des logisch-figuralen Schlussfolgerns sind die Verteilungen annähernd identisch.

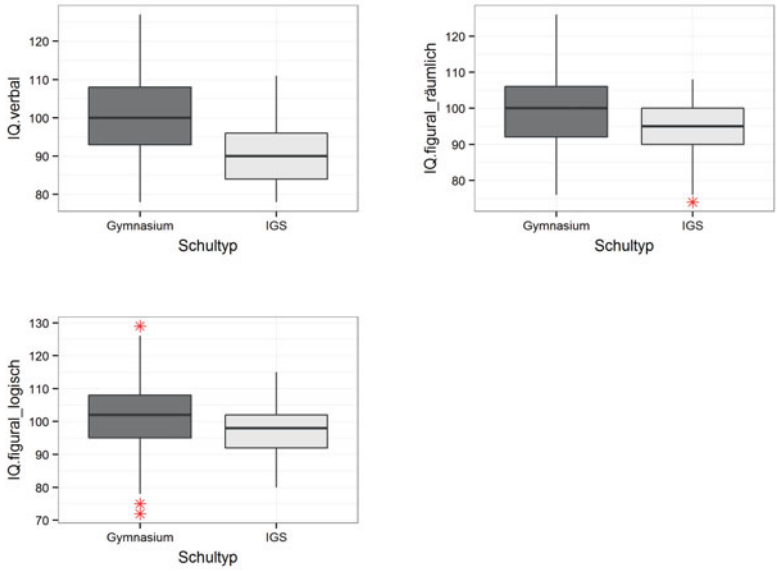


Abbildung 47: Boxplots kognitive Fähigkeiten je Bedingung

Ein Vergleich der Boxplots nach Schultyp und Verteilung zeigt, dass Gymnasiasten im Schnitt höhere Werte in den drei Aufgabenbereichen zur Messung der Intelligenz erzielten. Insgesamt zeigte sich eine höhere Streuung bei den Gymnasiasten als bei den Gesamtschülern. Bei den geringen Werten sind Gymnasiasten ebenso vertreten

wie Gesamtschüler. Die höchsten Werte sind durchweg bei den Gymnasiasten zu finden (vgl. Abbildung 48).

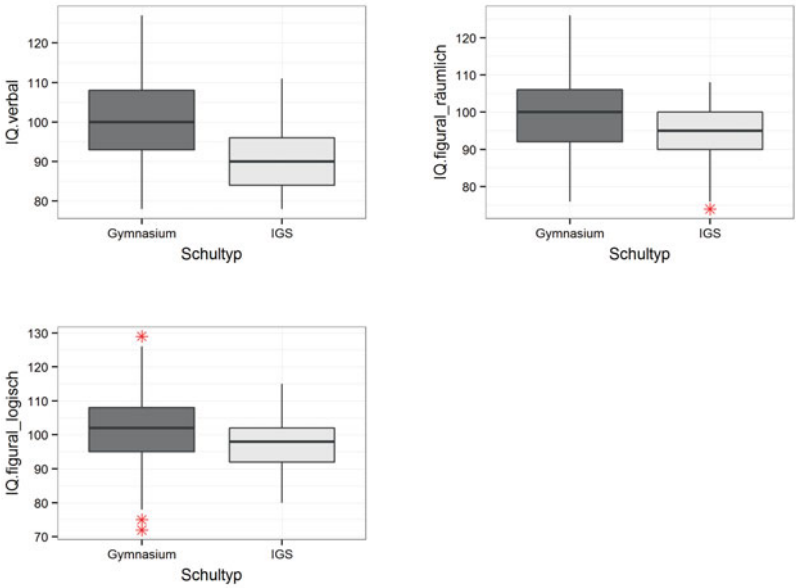


Abbildung 48: Boxplots kognitive Fähigkeiten je Schultyp

Zur Analyse der Zusammenhänge der drei Intelligenzkomponenten untereinander wurden Streudiagramme der drei gemessenen Intelligenzkomponenten (verbales Schlussfolgern, figural-räumliches Schlussfolgern und figural-logisches Schlussfolgern) gegeneinander erstellt. Die Diagramme wurden in Form von sogenannten „Jitter-Plots" erstellt (siehe Abbildungen 3 bis 5 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com).[31] Die drei Intelligenz-Komponenten lassen grafisch beim Blick auf die Jitter-Plots keine hohe Korrelation erkennen. Zur Prüfung dieser Schlussfolgerung wurde eine Produkt-Moment-Korrelation nach Bravais-Pearson (vgl. Eid et al., 2011, S. 506) berechnet. Die Ergebnisse be-

31 Bei Jitter-Plots werden zu den Punktkoordinaten Zufallszahlen addiert, damit eigentlich übereinanderliegende Punkte getrennt werden. Jede Erzeugung der Diagramme ergibt damit ein geringfügig anderes Bild.

legen eine schwache Korrelation zwischen den Bereichen des verbalen und des figural-räumlichen Schlussfolgerns, eine geringe jedoch signifikante Korrelation zwischen den Bereichen des verbalen und des figural-logischen Schlussfolgerns sowie eine signifikante Korrelation zwischen figural-räumlichem und figural-logischem Schlussfolgern (vgl. Tabelle 26).

Tabelle 26 Pearson-Korrelationen der drei Intelligenzskalen des I-S-T 2000 R ($N = 479$)

	1	2	3
1. Satzergänzung	-		
2. Würfelaufgaben	.08	-	
3. Matrizen	.22**	.26**	-

** Korrelationen ≥ .20 sind zweiseitig signifikant, $p < .01$

Die drei Korrelationen der Komponenten untereinander wurden insgesamt als eher niedrig eingeschätzt, so dass die Gefahr der Multikollinearität und damit verbundener instabiler Schätzungen der Gewichte des Regressionsmodells als gering zu bewerten ist.
Um für das spätere Modell jedoch die Anzahl an zu berücksichtigenden Variablen zu reduzieren, wurde die Zusammenfassung der Komponenten in Betracht gezogen. Eine Hauptkomponentenanalyse unter Verwendung der Varimax (orthogonale) Rotation erbrachte, dass eine Komponente extrahiert werden konnte, die 46% der Varianz aufklärt (vgl. Tabelle 27).

Tabelle 27 Eigenwerte und Anteil aufgeklärter Varianz Subskalen des I-S-T 2000 R

Komponente	Eigenwert	Anteil aufgeklärter Varianz
1	1.38	46.02 %
2	.92	30.64 %
3	.70	23.34 %

Basierend auf den Ergebnissen wurde davon abgesehen, statt der drei Intelligenzkomponenten eine oder zwei Linearkombinationen derselben zu verwenden, da der Anteil, der durch die Hauptkomponente beschriebenen Gesamtvarianz als nicht ausreichend angesehen wurde. Hinzu kommt, dass eine Zusammenfassung zu einer Hauptkomponente die Interpretierbarkeit des Modells erschwert.

2.3.6.2 Ergebnisse zu vorherigen Schulleistungen

Bezüglich der Noten in den Fächern Deutsch, Physik und Mathematik zeigen sich zwischen Treatment- und Kontrollgruppe kaum deskriptive Unterschiede. Im Schnitt hatten die Schüler die in Tabelle 28 aufgeführten Noten erreicht. Der Vergleich der Notenvergabe zwischen den Schultypen zeigt, dass Schüler der IGS im Schnitt schlechtere Bewertungen erhalten (vgl. Tabelle 28).

Tabelle 28 Mittelwert und Standardabweichung der Fachnoten nach Bedingung (und Schultyp)

Fachnote	TG[a] (*n* = 269)		KG[b] (*n* = 243)		Gymnasium (*n* = 423)		IGS[c] (*n* = 89)	
	M	*SD*	*M*	*SD*	M	SD	M	SD
Deutsch	2.72	0.05	2.80	0.06	2.68	0.86	3.15	0.75
Mathematik	2.82	0.06	2.82	0.06	2.73	0.98	3.27	0.86
Physik	2.95	0.06	2.83	0.07	2.77	1.01	3.46	1.21

[a]Treatmentgruppe, [b]Kontrollgruppe, [c]Integrierte Gesamtschule

Zur Analyse, ob die drei Noten wechselseitig korrelieren, wurden ebenso wie für die Subskalen des I-S-T 2000 R „Jitter-Plots “ (Deutsch, Mathematik, Physik) gegeneinander erstellt. Anhand der Plots (vgl. Abbildungen 6 bis 8, Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com) lässt sich ein statistischer Zusammenhang zwischen den Leistungen im Fach Mathematik und Deutsch sowie zwischen den Fächern Mathematik und Physik vermuten, welcher sich nicht auf Ausreißer zurückführen lässt. Zur Prüfung dieser Schlussfolgerung aus den Jitter-Plots wurde Spearmans Rangkorrelationskoeffizient berechnet, der sich für ordinalskalierte Daten eignet (vgl. Eid et al., 2011, S. 520). Die Ergebnisse weisen auf einen starken Zusammenhang zwischen der Deutsch- und der Mathematiknote sowie der Mathematik- und der Physiknote und einen mittleren Zusammenhang zwischen Deutsch- und Physiknote hin (vgl. Tabelle 29).

Tabelle 29 Rangkorrelationen nach Spearman der drei Fachnoten, (*N* = 512)

	1	2	3
1. Deutschnote	-		
2. Physiknote	.46**	-	
3. Mathematiknote	.52**	.50**	-

** p < .01, zweiseitig signifikant

Um für das spätere Modell die Anzahl an zu berücksichtigenden Variablen zu reduzieren, wurde ebenfalls die Zusammenfassung der Komponenten in Betracht gezogen. Eine Hauptkomponentenanalyse unter Verwendung der Varimax (orthogonale) Rotation erbrachte, dass eine Komponente extrahiert werden konnte, die 67 % der Varianz aufklärt (vgl. Tabelle 30).

Tabelle 30 Eigenwerte und Anteil aufgeklärter Varianz zu den drei Fachnoten, ($N = 479$)

Komponente	Eigenwert	Anteil aufgeklärter Varianz
1	2.01	67.04 %
2	.53	17.83 %
3	.45	15.14 %

Die Ergebnisse einer Parallelanalyse nach Horn (1965) bestätigen die Extraktion einer Komponente. So lag der zufällige Eigenwert bei 1.07.

Unter Berücksichtigung der deutlichen Korrelationen der Fachnoten und basierend auf den Ergebnissen der PCA wird in Betracht gezogen, statt der drei Fachnoten eine Linearkombination derselben zu verwenden.

2.3.6.3 Zusammenhang zwischen Schulleistungen und kognitiven Fähigkeiten

Abschließend wurden als nichtparametrisches Verfahren die Rangkorrelationen nach Spearman für die Noten und die erhobenen Subskalen des I-S-T 2000 R berechnet. Zur Wahl des Korrelationskoeffizienten bei Variablen mit unterschiedlichen Skalenniveau sei auf Eid et al. (2011, S. 538) hingewiesen. Die Ergebnisse weisen bis auf zwei Ausnahmen auf einen schwachen, jedoch signifikanten Zusammenhang zwischen Noten und den verwendeten Skalen des Intelligenztests auf (vgl. Tabelle 31). Wenig überraschend ist der geringe Zusammenhang zwischen Deutschnote und figural-räumlichem Schlussfolgern (Würfelaufgaben). Unerwartet ist jedoch, dass kein signifikanter Zusammenhang zwischen Physiknote und figural-räumlichem Schlussfolgern (Würfelaufgaben) gefunden wurde. Hier weicht der Wert auch von den Werten ab, welche die Autoren des Tests für die Korrelation zwischen Skalenwerten und Schulnoten berichten (vgl. Liepmann et al., 2007, S. 37).

Tabelle 31 Nichtparametrische Korrelationen der drei Fachnoten mit den verwendeten Subskalen des I-S-T 2000 R, ($N = 479$)

	Deutsch	Physik	Mathematik
Satzergänzung	-.30 **	-.25**	-.26**
Würfelaufgaben	-.07	-.09	-.18**
Matrizen	-.17**	-.17**	-.22**

**$p < .01$, zweiseitig signifikant

Basierend auf den eher niedrig einzustufenden Korrelationen der Intelligenzskalen mit den drei Fachnoten wird die Gefahr der Multikollinearität und damit verbundener instabiler Schätzungen der Gewichte des Regressionsmodells als gering eingeschätzt (vgl. Tabelle 31).

2.3.6.4 Diskussion der Analysen zu den Kovariaten

Möglicher Grund für die Abweichungen in den Itemkennwerten (Trennschärfe und Lösungswahrscheinlichkeit) und den Maßen für die interne Konsistenz könnten darauf zurückgeführt werden, dass es sich bei der Stichprobe um eine vergleichsweise junge Altersgruppe handelt. Auch wenn die Zuordnung der Messwerte zu den Standardwerten auf altersangepassten Normen basiert, wird vermutet, dass insbesondere die letzten und schwierigsten Aufgaben je Skala für die Stichprobe dieser Studie (Schüler der 7. und 8. Klasse) die Schüler im Alter von 12 bis 15 Jahren etwas überfordert haben könnten, da der Test in erster Linie auf die Erhebung der Intelligenz von Erwachsenen zielt.

Da davon auszugehen ist, dass die Schüler im Jugendalter etwas schlechter abschneiden als im späteren Erwachsenenalter, wird insgesamt vermutet, dass die untersuchte Gruppe eine höhere durchschnittliche Intelligenz aufweist als die Gesamtpopulation aller Schüler dieser Altersklasse. Dies ist aufgrund des hohen Anteils an Gymnasiasten (83%, 399 von 479 Schülern) erwartbar.

Die Analysen der Schulnoten ergaben ein ähnliches Bild: Gesamtschüler schnitten hier deutlich schlechter ab als Gymnasiasten. Sowohl in Bezug auf die Noten als auch auf die kognitiven Fähigkeiten zeigte sich eine höhere Streuung bei den Gymnasiasten als bei den Gesamtschülern. In Bezug auf die kognitiven Fähigkeiten waren bei den geringen Werten Gymnasiasten ebenso vertreten wie Gesamtschüler. Die höchsten Werte sind durchweg bei den Gymnasiasten zu finden.

Dieser Befund erscheint zunächst kontraintuitiv, da die Gesamtschule sich insbesondere an Schüler unterschiedlichster Leistungsbereiche richtet. Im Hin-

blick auf die Problematik der relativ frühen Selektion von Schüler im deutschen Schulsystem könnte das Ergebnis vor dem Hintergrund der oft diskutierten Bevorzugung von Kindern mit entsprechendem sozialen Hintergrund interpretiert werden (vgl. auch Ergebnisse der dritten PISA-Studie 2006, PISA-Konsortium Deutschland, 2007).

Für die Vergleichbarkeit von Treatment- und Kontrollgruppe ergeben sich hieraus keine negativen Konsequenzen. In der Folge sollte jedoch der Schultyp in der Auswertung unbedingt berücksichtigt werden und mögliche Interaktionen zwischen Schultyp und Bedingung (TG versus KG) – Interaktionen zwischen kognitiven Fähigkeiten und Bedingung sowie Interaktionen zwischen Schulnoten und Bedingung – analysiert werden.

Bezüglich des Zusammenhangs zwischen Noten und kognitiven Fähigkeiten ergaben sich weitestgehend erwartbare Ergebnisse. Überraschend war der geringe Zusammenhang zwischen Physiknote und figural-räumlichem Schlussfolgern (Würfelaufgaben). Ein möglicher Grund könnte darin bestehen, dass es sich bei der Physiknote in dieser Studie um einen Sonderfall handelt. So basiert diese Note entweder auf einem geringen Zeitabschnitt, weil der Physikunterricht erst in dem Schuljahr startete oder auf der Note im Fach Naturwissenschaften aus Klasse 6. Die Note aus Jahrgangsstufe 6 wurde in den Klassen, in denen noch keine Physiknote aus dem bisherigen Physikunterricht verfügbar war, ersatzweise herangezogen.[32]

Umgang mit dem Problem der Multikollinearität: Um instabile und ungenaue Schätzungen der Regressionskoeffizienten zu vermeiden, wird aufgrund der hohen Korrelationen der Fachnoten untereinander für die nachfolgenden Analysen das Ergebnis der PCA als Kovariate verwendet. In Bezug auf die Subskalen des I-S-T 2000 R und die Korrelationen zwischen Fachnoten und kognitiven Fähigkeiten werden die Ausgangswerte beibehalten, da aufgrund der schwachen wechselseitigen Korrelationen keine Multikollinearität besteht.

2.3.7 Einflussfaktoren bei der Anwendung des Lehrmaterials

2.3.7.1 Wahrgenommenes Lehrerengagement aus Schülersicht

Mögliche Einflussfaktoren bei der Anwendung des Lehrmaterials wurden aus zwei Perspektiven erhoben: erstens aus Sicht der Schüler und zweitens aus Sicht

32 Dies war der Fall, wenn der Physikunterricht zu Beginn der Studie erstmals startete und Strahlenoptik von der Lehrkraft als Einführungsthema gewählt wurde.

der Lehrkräfte. Im Rahmen des Motivationsfragebogens schätzten die Schüler in der Skala LES subjektiv das Engagement ihres Lehrers zu den drei Messzeitpunkten auf einer Likert-Skala von 1 (geringe Ausprägung des Engagements) bis 6 (hohe Ausprägung des Engagements) ein. Ein Blick auf die Mittelwerte (vgl. Tabelle 32) und Boxplots (vgl. Abbildung 49) zeigt, dass die Schüler in der Treatmentbedingung das Engagement ihres Lehrers zum ersten und zum dritten Messzeitpunkt, geringfügig höher bewerten als die Schüler der Kontrollbedingung.

Tabelle 32 Lehrerengagement aus Schülersicht je Bedingung zu den drei Messzeitpunkten

	Prätest			Posttest			Follow-up Test		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
TG[a]	243	4.56	1.18	238	4.15	1.39	255	4.36	1.28
KG[b]	222	4.41	1.28	213	4.14	1.33	216	4.19	1.40

Anmerkung. Skala von 1 (geringe Ausprägung) bis 6 (hohe Ausprägung)
[a]Treatmentgruppe,
[b]Kontrollgruppe

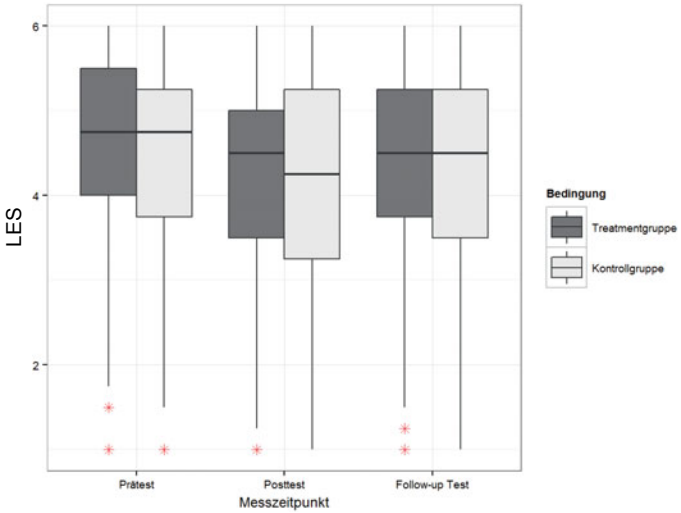


Abbildung 49: Boxplots des Lehrerengagements aus Schülersicht (LES) zu den drei Messzeitpunkten

Für die Durchführung der Studie besonders relevant waren mögliche Unterschiede zwischen Treatment- und Kontrollgruppe im Posttest (unmittelbar nach der Intervention). Die Schüler wurden explizit darum gebeten, ihre Einschätzung auf die letzten sechs Stunden zu beziehen. In den deskriptiven Werten zeigen sich (nahezu) identische Werte für die Mittelwerte und Standardabweichungen von $M = 4.15$ ($SD = 1.39$) bzw. $M = 4.14$ ($SD = 1.33$) in der Einschätzung der Schüler zwischen Treatment- und Kontrollgruppe nach der Intervention (post).

Die Analyse der Daten mittels eines Mehrebenenmodells (ohne Kovariaten) ergab zu keinem der drei Messzeitpunkte signifikante Unterschiede zwischen den Bedingungen (TG versus KG) – weder zu den einzelnen Messzeitpunkten noch in Bezug auf den Vergleich der Entwicklungsverläufe prä – post und prä – follow-up (vgl. Tabelle 33). So bestätigt sich auch in den Daten zum Anteil aufgeklärter Varianz mit Werten nahe 0, dass die Zuordnung zu den Bedingungen zur Erklärung der Varianz bedeutungslos ist. In beiden Bedingungen kam es jedoch zu einer signifikanten Abnahme der Einschätzung des Lehrerengagements im Schuljahresverlauf.

Tabelle 33 Ergebnisse des Mehrebenenmodells zur Analyse der Entwicklung des Lehrerengagements aus Schülersicht

Level (Stichprobengröße)						
Level 1 *(N = 1387)*	Messzeitpunkte					
Level 2 *(N = 470)*	Individuen					
Level 3 *(N= 21)*	Schulklassen					
Fixe Effekte						
Variable	*Erläuterung*		*b*	*(SE)*	$F_{(numDF, denDF)}$	*p*
Interzept	Durchschnittliche Einschätzung, Bedingung = KG		4.44	0.25	$312.68_{(1,860)}$	< .001
Veränderung (prä – post: t_2-t_1)	Durchschnittliche Veränderung, Bedingung = KG		-0.25	0.07	$11.81_{(1,860)}$	< .001
Veränderung (prä – follow-up: t_3-t_1)	Durchschnittliche Veränderung, Bedingung = KG		-0.20	0.07	$6.67_{(1,860)}$	= .01
Bedingung = TG	Unterschied TG und KG zu t_1		-0.13	0.35	$0.14_{(1,19)}$	= .709
Bedingung * Veränderung prä – post	Zusätzlicher Einfluss Bedingung = TG auf die Veränderung zwischen t_2 und t_1		-0.19	0.10	$3.63_{(1,860)}$	= .057
Bedingung * Veränderung prä – follow-up	Zusätzlicher Einfluss Bedingung = TG auf die Veränderung zwischen t_3 und t_1		0.01	0.11	$0.003_{(1,860)}$	= .956
Zufällige Effekte						
σ_u (Klasse)	0.77					
σ_p (Individuum)	0.66					
σ_ε (individuenspezifisch)	0.68					
ρ (Messzeitpunkt)	(ρ_{12}) 0.10	(ρ_{13}) -0.12	(ρ_{23}) 0.16			
g (Messzeitpunkt)	(g_1) 1.00	(g_2) 1.26	(g_3) 1.20			
Modellvergleich	*Berichtetes Modell*		*Leermodell*[a]			
Devianz	13883.94		3840.29			
df	14		11			
AIC	13941.94		3862.29			
BIC	14112.99		3919.88			
„Wirkung" des Treatments: Anteil aufgeklärter Varianz[a]						
R^2 Level 1	(t_1) 0.01	(t_2) < 0.01	(t_3) < 0.01			
R^2 Level 2	0.01					
R^2 Level 3	< 0.01					

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen. Das Modell sagt für jeden Schüler voraus, dass die Einschätzung des LES zum jeweiligen Messzeitpunkt der mittleren Einschätzung der Stichprobe entspricht. Man beachte, das Leermodell schätzt folgende 11 Parameter (*df* = 11): Interzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.

In den Verlaufsdiagrammen (vgl. Abbildung 50) erscheint der Rückgang des LES in der Treatmentgruppe bei den Gesamtschülern in der Postmessung größer als in der entsprechenden Kontrollgruppe, auch wenn diese Tendenz, wie die Mehrebenenanalyse zeigt (vgl. Tabelle 33), nicht signifikant ist.

Insgesamt bestätigen die Verlaufsdiagramme den leichten Abwärtstrend in der Bewertung der Schüler beider Bedingungen und Schularten im Schuljahresverlauf. Darüber weisen die Verlaufsdiagramme darauf hin, dass Gesamtschüler das Engagement ihres Lehrers etwas geringer bewerten als Gymnasiasten (vgl. Abbildung 50). Aufschluss für mögliche Gründe dieser Entwicklung könnte die Analyse des Lehrernotizheftes ergeben.

2.3.7.2 Auswertung des Lehrernotizhefts

Um die Perspektive der Lehrkräfte bei der Umsetzung der Intervention zu erfassen wurden die Lehrkräfte gebeten, jede Unterrichtsstunde der Intervention in einem Lehrernotizheft zu bewerten und besondere Vorkommnisse oder Störungen zu notieren. Die vorliegenden Ergebnisse sind in Tabelle 34 zusammengefasst. Die Bewertung umfasste drei Skalen: erstens die Bewertung der eingesetzten Aufgaben, zweitens die Einschätzung des Unterrichtsablaufs und drittens die Bewertung der Lernbilanz.
Der Mittelwert zur Einschätzung der Aufgabenschwierigkeit und des Aufgabenumfangs liegt bei $M = 1.69$ (0.76) bzw. $M = 2.35$ (0.77) in der Treatmentgruppe und bei $M = 1.53$ (0.73) bzw. $M = 2.02$ (0.86) in der Kontrollgruppe. Damit wurden die Aufgaben im Bereich „nicht sehr schwierig“ bis „schwierig aber lösbar“ bzw. „umfangreich, aber in einer Stunde lösbar“ mit der Tendenz zur Einschätzung eines zu hohen Aufgabenumfangs in der Treatmentgruppe bewertet. Die Beurteilerübereinstimmung je Bedingung (Treatment versus Kontrollgruppe) kann für die Bewertung der Aufgaben und des Unterrichtsablaufs als zufriedenstellend ($ICC = .51 – .57$) bis gut ($ICC = .61 – .90$) bewertet werden (vgl. Cichetti & Sparrow, 1981).[33] Die niedrigeren Übereinstimmungen bezüglich des Ausmaßes an kognitiver Aktivierung (Lernbilanz, Frage 3) spiegeln wieder, dass die Meinungen über das Maß an kognitiver Aktivierung in der Kontrollbedingung unterschiedlich bewertet wird, was auf die gewollte Variation im Ausmaß der kognitiven Bedingungen zurückzuführen ist.

33 Da es sich um ordinalskalierte Daten handelt, wurde als Maß für die Beurteilerübereinstimmung Kendalls Konkordanzkoeffizient *W* berechnet und der anschaulichere *ICC*, der jedoch unter dem Vorbehalt zu interpretieren ist, dass die Voraussetzung des Intervallskalenniveaus im strengen Sinn nicht erfüllt ist.

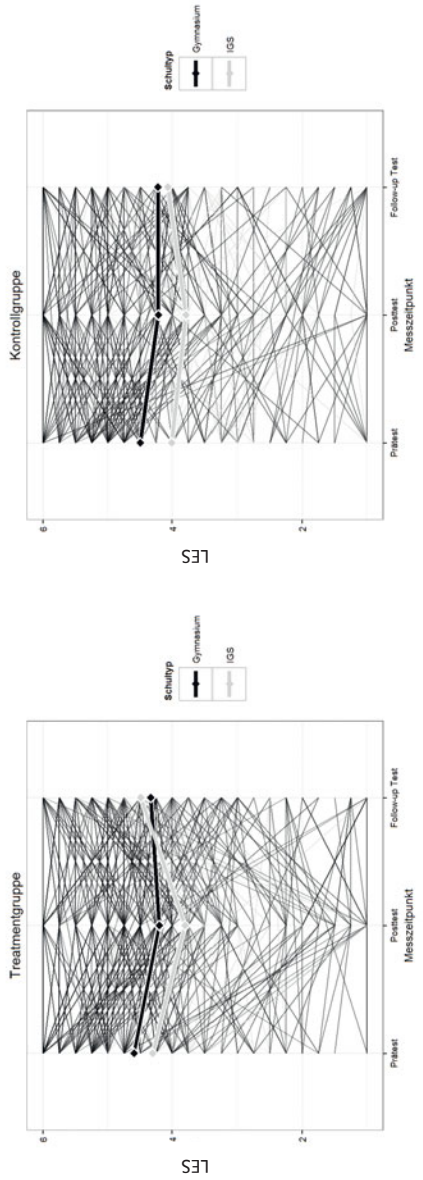


Abbildung 50: Verlaufsdiagramme des Lehrerengagements aus Schülersicht (LES) zu den drei Messzeitpunkten

Tabelle 34 Einschätzung der Unterrichtsqualität der $N = 10$ teilnehmenden Lehrkräfte gemittelt über die sechs geplanten Unterrichtsstunden

Item	TG				KG			
	M	*SD*	*Modus*	$ICC_{unjustiert}$	*M*	*SD*	*Modus*	$ICC_{unjustiert}$
Bewertung der Aufgaben								
Aufgabenschwierigkeit	1.69	0.76	1	.62	1.53	0.73	1	.51
Aufgabenumfang	2.35	0.77	3	.52	2.02	0.86	2	.42
Kendalls *W*	.50	$\chi^2_{(1)} = 4.50$, $p = .034$			.40	$\chi^2_{(1)} = 3.57$, $p = .059$, n.s.		
Skala 0 - 3,	0 - zu einfach / zu wenig, 1 - nicht sehr schwierig / gut umsetzbar, 2 - schwierig / umfangreich, aber lösbar, 3 - zu schwierig / zu umfangreich							
Unterrichtsablauf								
Der Unterricht hat pünktlich begonnen	3.10	0.85	3	.65	3.11	0.75	3	.20
Den Schülern war jederzeit klar, was sie tun sollten	2.92	0.85	3	.61	3.13	0.82	3	.44
Die gesamte Unterrichtsstunde wurde für den Lernstoff verwendet	3.38	0.82	4	.56	3.39	0.70	4	.27
Kendalls *W*	.24	$\chi^2_{(1)} = 4.22$, $p = 1.21$, n.s.			.04	$\chi^2_{(1)} = 0.77$, $p = 2.06$, n.s.		
Lernbilanz								
Ich schätze: Die Schüler haben in dieser Stunde etwas dazu gelernt	3.16	0.67	3	.61	3.19	0.68	3	.61
Die Lernziele dieser Unterrichtsstunde wurden erreicht	2.91	0.93	3	.90	2.85	0.94	3	.57
Ich schätze: Der Unterricht hat die Schüler zum Nachdenken angeregt	2.93	0.77	3	.67	3.04	0.75	3	.32
Kendalls *W*	.44	$\chi^2_{(1)} = 4.50$, $p = .018$			.45	$\chi^2_{(1)} = 8.06$, $p = .018$		
Skala 1-4	1 - stimme nicht, 2 - stimme eher nicht zu, 3 -stimme eher zu, 4 - stimme zu							

Bezüglich des Unterrichtsablaufs divergieren die Bewertungen. Unter Berücksichtigung der Auswertung der Kommentare der Lehrer zu den einzelnen Stunden lässt sich dieses Ergebnis auf Gegebenheiten im Schulalltag zurückzuführen. Die genannten Störungen (z.B. Raumverlegungen, Besprechung von Klassenfahrten u.a.) stellen jedoch keine wesentlichen Einschränkungen für die Vergleichbarkeit der beiden Bedingungen dar, da die Lernzeit in beiden Bedingungen parallel gehalten wurde.

Der Mittelwert der Zustimmung der Fragen zur Unterrichtsqualität rangiert zwischen $M = 2.92$ (0.85) bis $M = 3.38$ (0.82) in der Treatment- und $M = 2.85$ (0.94) bis $M = 3.11$ (0.75) in der Kontrollgruppe und liegt damit im Bereich zwischen „stimme eher zu“ (3) und „stimme zu“ (4).

Ein Mann-Whitney-U-Test für unabhängige Stichproben (hier Treatment- versus Kontrollgruppe) bei ordinalskalierten Daten ergab mit Ausnahme der Einschätzung des Aufgabenumfangs keine signifikanten Unterschiede in den Einschätzungen seitens der Lehrkräfte zwischen Treatment- und Kontrollgruppe.

Während die Aufgabenschwierigkeit zwischen nicht „sehr schwierig“ oder „schwierig aber machbar“ eingestuft wurde, schätzten die Lehrkräfte den Aufgabenumfang in der Treatmentbedingung als signifikant höher ein als in der Kontrollbedingung ($U(10,10) = 1015.00$, $z = 2.05$ $p = .040$). Um diesem Befund nachzugehen, wurde die Einschätzung des Aufgabenumfangs für die einzelnen Unterrichtsstunden betrachtet.

Ein Blick auf die einzelnen Stunden zeigt, dass der Umfang für alle Unterrichtsstunden außer der dritten nicht signifikant unterschiedlich bewertet wurde. So lag die Einschätzung des Aufgabenumgangs der Stunden 1 bis 2 und 4 bis 6 im Schnitt bei 2 „Die Aufgaben sind umfangreich, aber in einer Stunde lösbar“; außer für die dritte Unterrichtsstunde, in welcher die Aufgaben als zu umfangreich eingeschätzt wurden.

Hier weichen die Ergebnisse für die Treatment- und die Kontrollgruppe auch signifikant ab: ($U(10,10) = 10.00$, $z = 2.56$, $p = .021$).

Mit Blick auf den Unterrichtsablauf lässt sich dieser Befund leicht erklären. So wurde in der dritten Unterrichtsstunde die Strahlenkonstruktion behandelt. Zwecks Datenerhebung wurden hier die Arbeitsblätter der Schüler eingesammelt und zwar ausschließlich in der Treatmentgruppe. Die Kontrollgruppe hielt parallel einen Lehrervortrag und übte die Strahlenkonstruktion später. Die Einschätzung der Lehrkräfte deckt sich hierbei mit der Unterrichtsbeobachtung. So war das Ein- und Austeilen der Blätter mit einem deutlich höheren Zeitaufwand verbunden als in der Kontrollgruppe. Zudem musste in der Treatmentgruppe zusätzliche Zeit aufgewandt werden, um den Schülern den Zweck des Einsammelns zu erklären.

Die Analyse der übrigen Items erbrachte keine Unterschiede zwischen Treatment- und Kontrollgruppe. Dies ist ebenso wenig für die dritte Unterrichtsstunde als auch für die übrigen Stunden der Fall. So wurde weder der Lernzuwachs noch das Erreichen der Lernziele für die beiden Gruppen unterschiedlich bewertet.

In einer offenen Frage wurden die teilnehmenden Lehrkräfte zudem gebeten Anmerkungen, besondere Vorkommisse und offene Punkte zu nennen. Die Antworten der Lehrer lassen sich wie folgt zusammenfassen:

- Kritik wurde an der Thematisierung des virtuellen Bildes geübt. Hier wurde vor allem mehr Zeit für die Einführung und Besprechung des virtuellen Bildes gefordert.

- Das Austeilen und Einsammeln der Arbeitsblätter zwecks Datenerhebung wurde als zeitaufwendig eingeschätzt.
- Seitens der teilnehmenden Gesamtschullehrer wurde angemerkt, dass mehr Möglichkeiten der Binnendifferenzierung zwischen schnellen und langsamen Schülern wünschenswert gewesen wären.
- Bezüglich der Arbeitsblätter ergab sich in manchen Klassen das Problem, dass die Schüler die Arbeitsaufträge nicht genau lasen und es hierdurch zu zeitlichen Verzögerungen kam.
- Positiv wurde genannt, dass die Arbeitsblätter den Schülern einen Überblick verschafften, die sechste Unterrichtsstunde Gelegenheit zur Wiederholung und Vertiefung bot und dass die Schüler die Aufgabenblätter zu Teilen mit Interesse und weitgehend selbstständig erarbeiten konnten.

2.3.7.3 Diskussion zu den Einflussfaktoren bei der Anwendung des Lehrmaterials

Das Mehrebenenmodell zur Entwicklung des Lehrerengagements aus Schülersicht belegt, dass keine signifikanten Unterschiede in der Einschätzung des LES bestehen. Insbesondere zum zweiten Messzeitpunkt ergaben sich nahezu gleiche Werte in der Einschätzung der Schüler (siehe arithmetisches Mittel und Standardabweichung des LES), daher kann von einem vergleichbar wahrgenommenen Engagement der Lehrkräfte in den Bedingungen ausgegangen werden. Zur Berücksichtigung von Unterschieden zwischen den Schulen, welche sich insbesondere durch die niedrigeren Werte der Gesamtschulen zeigen, wird das LES vor der Intervention (erster Messzeitpunkt) als Kovariate in den folgenden Analysen einbezogen.

Auch in den Analysen des Lehrernotizhefts zeigen sich auf Schul- und Klassenebene keine Unterschiede, welche die Vergleichbarkeit zwischen Treatment- und Kontrollgruppe einschränken würden.

Die genannten Kritikpunkte ergeben sich vorwiegend aus Erfordernissen der empirischen Erhebung. So zielte das erstellte Unterrichtsmaterial darauf, bis auf den intendierten Unterschied zwischen Treatment- und Kontrollgruppe vergleichbare Bedingungen herzustellen. Entsprechend wurde weniger Wert auf die Möglichkeit der Leistungsdifferenzierung gelegt. Aspekte, welche für die Lerninhalte mit geringerer Priorität bewertet wurden, wie das virtuelle Bild, konnten durch den vorgegebenen Zeitraum weniger ausführlich behandelt werden.

Das Einsammeln der Arbeitsblätter diente dazu, Einblicke in den Lernprozess zu erhalten. Gegebenenfalls hätte hier eine andere Methode der Datenerhebung zu einem geringeren Aufwand für die teilnehmenden Lehrer und Schüler geführt.

Letzterer Punkt könnte eventuell erklären, warum im Verlauf Schüler beider Bedingungen das Engagement ihres Lehrers – bezogen auf den Zeitraum der Intervention geringer bewerten als zuvor oder danach. Möglicherweise lässt sich dieser Effekt auf eine gewisse „Testmüdigkeit" seitens der teilnehmenden Schüler und Lehrkräfte zurückführen.

2.3.8 Einblick in die Lernprozesse beim Umgang mit Repräsentationen

2.3.8.1 Auswertung der Arbeitsblätter 4 und 7 zur Bildkonstruktion

Um Einblicke in den Lernprozess zu erhalten, wurde eine geschichtete Zufallsauswahl an Arbeitsblättern, die im Rahmen der Intervention eingesammelt wurden, ausgewertet. Sowohl in der Treatment- als auch in der Kontrollbedingung wurde das Aufgabenblatt 7 „Wiederholung zur Bildkonstruktion" siehe Anhang C1 auf der Produktseite zu diesem Buch unter www.springer.com, eingesammelt, welches die Schüler in der fünften Unterrichtsstunde bearbeiteten. Folgende Fragen wurden im Rahmen der Auswertung von Aufgabenblatt 7 untersucht:

1. Unterscheiden sich Schüler in den beiden Bedingungen (TG versus KG) signifikant hinsichtlich des Erfolgs beim Bearbeiten der Aufgaben?
2. Sind Leistungsunterschiede, die im Leistungspost- und im Konzeptposttest gemessen wurden, bereits im Lernprozess zu erkennen:
   - Unterscheiden sich die Schüler in unterschiedlichen Leistungsbereichen, hohes Leistungsniveau, mittleres Leistungsniveau und unteres Leistungsniveau bereits bei der Bearbeitung des Aufgabenblatts?
   - Korreliert der Erfolg beim Bearbeiten der Aufgaben mit den jeweiligen Posttestergebnissen signifikant? Wenn ja, wären bezüglich der Auswertung des Aufgabenblatts 7 für die A1 (Kohärenzaufgabe, vgl. auch Treatment von Scheid, 2013) höhere Korrelationen mit dem Leistungstest und für die Aufgaben A2 und A3 (Linse < Schirm, Abdeckaufgabe) höhere Korrelationen mit dem Konzeptposttest zu erwarten.

In der Treatmentgruppe wurde zudem das Aufgabenblatt 4 „Die Konstruktion der reellen Bildfälle" aus der dritten Unterrichtsstunde ausgewertet (siehe Anhang C1 auf der Produktseite zu diesem Buch unter www.springer.com). Folgende Fragen wurden hierzu untersucht:

1. Besteht ein Zusammenhang zwischen Lernprozessen zum Verständnis und Umgang der zentralen schematischen Repräsentation des Strahlengangs mit dem Leistungspost- bzw. dem Konzeptposttest?
2. Sofern der Erfolg beim Bearbeiten der Aufgaben mit den jeweiligen Testergebnissen signifikant korreliert, wären bezüglich der Auswertung des Aufgabenblattes 4 höhere Korrelationen mit dem Leistungs- als mit dem Konzeptposttest zu erwarten.

Die geschichtete Zufallsauswahl wurde nach folgenden Kriterien gezogen: pro Bedingung Treatment- versus Kontrollgruppe wurden jeweils fünf Schüler des unteren Quartils des Konzeptpost- und des Leistungsposttests, fünf Schüler des oberen Quartils des Konzeptpost- und Leistungsposttests sowie je fünf Schüler der mittleren Prozenträge des Konzeptpost- und des Leistungsposttest per Zufall (Zufallsstichprobe in SPSS) ausgewählt (vgl. Tabelle 35 und Tabelle 36). Weitere Informaitionen zur Verteilung von Jungen und Mädchen bzw. den Anteilen an Gymnasiasten und Gesamtschülern finden sich in den Tabellen 18 und 19 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com.

Tabelle 35 Kreuztabelle zur Verteilung der ausgewählten Schüler je Test (abhängige Variable) und Bedingung

		Abhängige Variable		
		LT[a]	KT[b]	∑
Bedingung	TG[c]	15	15	30
	KG[d]	15	15	30
∑		30	30	60

[a]Leistungstest, [b]Konzepttest, [c]Treatmentgruppe, [d]Kontrollgruppe

Die Auswertung des Aufgabenblatts 7 erfolgte nach folgendem Schema: je Aufgabe konnten sowohl in der Treatment- als auch in der Kontrollbedingung maximal 4 Punkte erzielt werden. Daraus ergibt sich für das Aufgabenblatt 7 eine Höchstpunktzahl von 3 x 4 Punkten = 12 Punkten. Die Punktevergabe erfolgte nach einem festgelegten Schema, bei dem sowohl für die Zeichnung als auch für die Begründung je 2 Teilpunkte (TP) erzielt werden konnte. Die Vergabe der Punkte wurde von einem geschulten Auswerter mit fachdidaktischem Hintergrund vor-

genommen und von der Autorin im Hinblick auf die Einhaltung des Auswertungsschemas geprüft.

Tabelle 36 Kreuztabelle zur Verteilung der ausgewählten Schüler je Test (abhängige Variable) und Quartil des Leistungsbereichs Physikleistung und konzeptuelles Verständnis

		Abhängige Variable		
		LT[a]	KT[b]	$\sum$
Quartil	$Q_{0.25}$	10	10	20
	$Q_{0.25} - Q_{0.75}$	10	10	20
	$Q_{0.75}$	10	10	20
$\sum$		30	30	60

[a]Leistungstest, [b]Konzepttest

Tabelle 37 Deskriptive Statistiken zur Auswertung von A1-A3 des Aufgabenblatt 7 „Wiederholung zur Bildkonstruktion“, ($N = 60$)

	A1	A2	A3	Gesamt
	2 TP Zeichnung, 2 TP Begründung	2 x 1 TP Zeichnung, 2 x 1 TP Begründung	2 x 1 TP Zeichnung, 2 x 1 TP Begründung	3 x 4 Punkte
M	1.61	1.69	1.00	4.05
(SD)	(1.46)	(1.40)	(1.26)	(3.39)
Range	0 - 4	0 - 4	0 - 4	0 - 12
α (A1-A3, 3 Items)	.74			

Anmerkung. A1: 2 x 2TP; A2 und A3 je 4 x 1TP

Wie an den deskriptiven Statistiken ersichtlich ist, ist die Bearbeitung der Aufgaben 2 und 3 durch einen starken Bodeneffekt gekennzeichnet (vgl. Tabelle 37). Insbesondere die Abdeckaufgabe A3 fällt durch einen niedrigen Mittelwert auf. Möglicherweise bereitet die weitverbreitete Schülervorstellung, die Linse schneide das Bild ab, Schülern Probleme im Umgang mit der Strahlenkonstruktion und der verbalen Begründung ihrer Schlussfolgerungen aus der Zeichnung. Die interne Konsistenz der Aufgaben als Schätzer für die Reliabilität kann mit $\alpha = .75$ als zufriedenstellend eingestuft werden.

Zur Untersuchung der ersten Frage wurden die Mittelwerte der Aufgaben mittels eines t-Tests für unabhängige Stichproben verglichen (siehe Tabelle 38).

In der Analyse zeigten sich keine signifikanten Mittelwertunterschiede zwischen Schülern in den beiden Bedingungen (TG versus KG).

Tabelle 38 t-Test zur statistischen Prüfung von Mittelwertunterschieden zwischen TG und KG bei der Bearbeitung des Aufgabenblatts 7 „Wiederholung zur Bildkonstruktion", ($N = 60$)

	Gesamt	A1	A2	A3
TG - *M (SD)*	3.40 (2,70)	1.37 (0.96)	1.48 (0.95)	1.03 (0.96)
KG - *M (SD)*	4.70 (3.90)	1.87 (1.50)	1.90 (1.72)	0.97 (1.52)
t (58)	1.50, $p = 1.89$	1.33, $p = 2.52$	1.16, $p = .84$	0.20, $p = 1.40$
$M_{Diff.}$ ($SD_{Diff.}$)	1.30 (0.86)	0.50 (0.38)	0.42 (0.36)	0.67 (0.33)

Im Hinblick auf den Erfolg in der Aufgabenbearbeitung sind signifikante Unterschiede von Schülern im unteren, mittleren und oberen Leistungsniveau des Leistungspost- bzw. Konzeptposttests ersichtlich (Forschungsfrage 2a).
Mittels einer univariaten Varianzanalyse wurde geprüft, ob signifikante Unterschiede zwischen den Schülern der unterschiedlichen Leistungsbereiche bestanden. Hierbei bildeten die Faktorstufen die Zuteilung zu den drei Leistungsbereichen und die abhängige Variable die Gesamtpunktzahl in den drei Aufgaben. Die Analyse bestätigt, dass signifikante Unterschiede, die in den Posttests festgestellt wurden, sich bereits bei der Aufgabenbearbeitung zeigen (F (2,57) = 8.61, $p < .001$, $\eta_p^2 = .23$, $\omega^2 = 0.20$, großer Effekt).

Zudem wurde der Effekt jeder Faktorstufe paarweise mit dem Effekt der vorherigen Faktorstufe verglichen (Post-hoc-Test Tukey). Die Ergebnisse zeigen (vgl. Tabelle 39), dass Schüler im unteren Leistungsposttestniveau die Aufgaben weniger erfolgreich bearbeiteten als Schüler des mittleren und oberen Niveaus, während sich Schüler des mittleren und oberen Leistungsniveaus nicht signifikant unterschieden.

Die Analyse der Korrelationen des Erfolgs beim Bearbeiten der Aufgaben mit den jeweiligen Posttestergebnissen (Forschungsfrage 2b) zeigte signifikante Zusammenhänge zwischen der Bearbeitung der Teilaufgaben A1 bis A3 sowie der Gesamtwertung des Arbeitsblatts und der Gesamtpunktzahl im Konzeptpostest, nicht jedoch zwischen der Bearbeitung des Aufgabenblatts und der Gesamtpunktzahl im Leistungsposttest.

Tabelle 39 Paarweiser Vergleich der Schüler der verschiedenen Leistungsniveaus bezüglich der Gesamtpunktzahl für die Bearbeitung des Aufgabenblatts 7, ($N = 60$)

Quartil	*n*	Vergleich	*M*	*SD*	*p*
Q_{25}	20	Mittleres gegen unteres Niveau	2.45	0.96	= .034
$Q_{0.25} - Q_{0.75}$	20	Oberes gegen mittleres Niveau	1.47	0.96	= .279
Q_{75}	20	Oberes gegen unteres Niveau	3.93	0.96	< .001

Die höchste Korrelation findet sich zwischen der Gesamtwertung des Arbeitsblatts und der Gesamtpunktzahl im Konzeptpostest, $r(58) = .41$, $p < .001$. Entgegen der Vorannahme besteht zwar ein signifikanter Zusammenhang zwischen der Punktzahl für Aufgabe 1 (Kohärenzaufgabe) und dem Ergebnis des Konzeptposttests, nicht jedoch zwischen der Punktzahl für Aufgabe 1 und dem Leistungsposttest. Die Punktzahlen für A2 und A3 korrelieren, wie erwartet, höher miteinander als die jeweilige Korrelation mit der Punktzahl für Aufgabe 1 (vgl. Tabelle 40).

Tabelle 40 Produkt-Momentkorrelation zwischen den Posttestergebnissen und den Arbeitsaufgaben von Arbeitsblatt 7 *($N = 40$; $n_{Q25} = 10$, $n_{Q25\text{-}Q75} = 20$, $n_{Q75} = 10$)*

	Test / Aufgabe (A)	1	2	3	4	5	6
1	Leistungsposttest	-					
2	Konzeptposttest	.35*	-				
3	A_1	.15	.33*	-			
4	A_2	.06	.34*	.39*	-		
5	A_3	.08	.36*	.36*	.63**	-	
6	$A_{1\text{-}3}$ Gesamt	.09	.41**	.74**	.83**	.81**	-

*p < .05, **p < .001, zweiseitig signifikant

Zwischen der Gesamtwertung des Aufgabenblatts und den Subskalen des I-S-T 2000 R, sowie zwischen der Gesamtwertung und den Fachnoten in Mathematik, Physik und Deutsch konnten in der Teilstichprobe von ($n = 40$) keine signifikanten Korrelationen festgestellt werden. Die relativ stärksten Korrelationen bestanden hier zwischen Physiknote und Gesamtwertung des Aufgabenblatts 7 ($r(38) = -.30$, *n.s.*) sowie zwischen Würfelaufgaben (räumlich-figurales Schlussfolgern) und Gesamtwertung des Aufgabenblatts ($r(38) = .23$, *n.s.*).

Im Zug der Auswertung von Aufgabenblatt 4 wurde untersucht, ob ein Zusammenhang beim Bearbeiten von Aufgabenblatt 4 (bildhaftes, verbales und

mathematisches Verständnis der Strahlenkonstruktion) und dem Lernerfolg im Leistungs- bzw. Konzeptposttest besteht.

Die Auswertung des Aufgabenblatts 4 erfolgte analog zur Auswertung von Aufgabenblatt 7: je Teilaufgabe (4a, b, c) konnten maximal 8 Punkte erzielt werden; daraus ergibt sich für das Aufgabenblatt 4 eine Höchstpunktzahl von 3 x 8 Punkten = 24 Punkten.

Tabelle 41 Deskriptive Statistiken zur Auswertung von 4a-4c des Aufgabenblatts 4 „Die Konstruktion der realen Bildfälle“ ($N = 30$)

Aufgabe	4a	4b	4c	Gesamt
Teilpunkte (TP)	4 TP Zeichnung 2 TP Begründung 2 TP Messen	4 TP Zeichnung 2 TP Begründung 2 TP Messen	4 TP Zeichnung, 2 TP Begründung 2 TP Messen	3 x 8 Punkte
M	4.78	5.91	5.73	16.43
(SD)	(2.55)	(2.25)	(2.23)	(6.63)
Range	0 - 8	1 - 8	0 - 8	1 - 23.50
α (4a - 4c, 3 Items)		.94		

Wie an den deskriptiven Statistiken ersichtlich ist (vgl. Tabelle 41), findet ein deutlicher Lerneffekt zwischen der Bearbeitung von 4a zu 4b statt (dreigliedrige Strahlenkonstruktion) statt. Die leichte Abwärtstendenz in der Gesamtwertung von Teilaufgabe 4c kann auf die erhöhte Anforderung zurückgeführt werden, die Größen- und Streckenverhältnisse verbal und mathematisch frei zu formulieren. In Teilaufgabe 4b bestanden hier noch mehr Unterstützungsangebote (repräsentationales Fading-out). Die interne Konsistenz der Aufgaben als Schätzer für die Reliabilität kann mit $\alpha \geq .90$ als sehr gut bewertet werden.

Wie erwartet, ergab die Analyse der Korrelationen des Erfolgs beim Bearbeiten der Aufgaben mit den jeweiligen Posttestergebnissen tendenziell stärkere Zusammenhänge zwischen der Bearbeitung der Teilaufgaben 4a-c sowie der Gesamtwertung des Arbeitsblatts und der Gesamtpunktzahl im Leistungsposttest als zwischen der Bearbeitung der Aufgaben und der Gesamtpunktzahl im Konzeptposttest (vgl. Tabelle 42).

Zwischen der Gesamtwertung des Aufgabenblatts und den Subskalen des I-S-T 2000 R, sowie zwischen der Gesamtwertung und den Fachnoten in Mathematik, Physik und Deutsch konnte in der Teilstichprobe von (*n = 30*) lediglich eine signifikante Korrelation zwischen Physiknote und Gesamtwertung des Aufgabenblatts 4 ($r(28) = -.41, p < .05$) festgestellt werden.

Tabelle 42 Produkt-Momentkorrelation zwischen den Posttestergebnissen und den Arbeitsaufgaben von Arbeitsblatt 4, (n = 30)

	Testwert / Aufgabe	1	2	3	4	5	6
1	Leistungsposttest	-					
2	Konzeptposttest	.35*	-				
3	Aufgabe 4a	.57**	.66**	-			
4	Aufgabe 4b	.62**	.51**	.82**	-		
5	Aufgabe 4c	.70**	.54**	.85**	.83**	-	
6	$4_{a\text{-}c}$ Gesamt	.66**	.61**	.95**	.94**	.95**	-

Anmerkung: Arbeitsblatt 4 wurde in dieser Form nur von der Treatmentgruppe bearbeitet
* $p < .05$, ** $p < .001$, zweiseitig signifikant

2.3.8.2 Diskussion der Ergebnisse zur Analyse der Lernprozesse

Die Auswertung der Arbeitsblätter bestätigt, dass der Umgang mit der zentralen Repräsentation der Strahlenkonstruktion bereits in den Lernprozessen entscheidend für den Lernerfolg ist. Die Frage, ob der erfolgreiche Umgang mit zentralen Repräsentationen ein geeigneter Prädiktor für den Lernerfolg unmittelbar nach dem Unterricht darstellt, kann somit mit „ja“ beantwortet werden. Die guten bis sehr guten Schätzwerte für die Reliabilität sprechen für die Zuverlässigkeit der Datenbasis.

Insgesamt kann die Bearbeitung der Aufgaben als mittelstarker Prädiktor für die Lernleistung im Konzept- und Leistungsposttest gewertet werden. Die signifikanten Korrelationen im mittleren Bereich von Aufgabenbearbeitung und der Physiknote stützen diesen Befund, zumal keine signifikanten Korrelationen mit den anderen Fachnoten bestanden.

Da sich die Analysen wegen des hohen Auswertungsaufwands auf eine kleine Stichprobe von 40 Schülern beziehen, könnte der tatsächliche Zusammenhang zwischen Aufgabenbearbeitung und Lernleistung in der Gesamtstichprobe sogar noch stärker sein. Erwartungsgemäß waren Aufgaben zur Bearbeitung der Strahlenkonstruktion bessere Prädiktoren zur Vorhersage des Leistungspostest-Ergebnisses und Aufgaben, welche Schülervorstellungen (Abdeckaufgaben) thematisierten, bessere Prädiktoren für die Ergebnisse im Konzeptposttest.

Unterschiede in der Bearbeitung zwischen den Bedingungen (TG versus KG) konnten zu dem Zeitpunkt des Einsammelns der Aufgabenblätter (gegen Mitte bzw. zu Beginn des letzten Drittels der Intervention) nicht festgestellt werden.

2.3.9 Ergebnisse zu den untersuchten Hypothesen

2.3.9.1 Methodisches Vorgehen: Mehrebenenanalyse

Zur Analyse der Fragestellungen wurde eine Mehrebenenanalyse gewählt. Hierbei handelt es sich prinzipiell um ein Regressionsmodell, welches aber im Gegensatz zu klassischen Regressionsmodellen in der Lage ist, die Mehrebenenstruktur der Daten zu berücksichtigen (vgl. Eid et al., 2011, S. 699). Eine solche Mehrebenenstruktur zeichnet sich dadurch aus, dass die Elemente auf den unteren Ebenen in jeweils genau ein Elemente der nächst höheren Ordnung eingebettet sind (vgl. Ditton, 1998, S. 11; Eid et al., 2011, S. 701). Mehrebenenmodelle werden daher auch als hierarchische lineare Modelle bezeichnet (Bryk & Raudenbush, 1992; Snijders & Bosker, 1999; Hox, 2002): Werden solche hierarchisch strukturierten Datensätze wie einfache Zufallsstichproben behandelt, besteht das Risiko falsche Schlussfolgerungen bei der Interpretation von Zusammenhängen zu ziehen (vgl. Eid et al., 2011, S. 700). Die eigentlich notwendige Unabhängigkeit der Beobachtungen ist in hierarchischen Datenstrukturen nicht erfüllt, da die untersuchten Individuen innerhalb der Gruppen oder Aggregateinheiten gemeinsamen Einflüssen oder Erfahrungen unterliegen, welche für die Einheiten eines Aggregats charakteristisch sind.

Bei der hier vorliegenden Stichprobe handelt es sich um eine solche mehrstufige Stichprobenziehung. So wurden die Elemente nicht unabhängig voneinander ausgewählt, sondern eine Zufallsauswahl von Schulen getroffen, welche wiederum eine Auswahl von Klassen und eine Auswahl von Schülern enthielt.

Die hier vorliegende Struktur stellt sich wie folgt dar: Pro Individuum wurde die fachliche Leistung, das konzeptuelle Verständnis und die Motivation zu drei Messzeitpunkten erhoben. Jedes Individuum war einer der 21 Klassen zugeordnet (mit einer Klassenstärke, welche zwischen 17 bis 28 Schüler variierte). Die Schüler wurden als Klassen der Treatment- und Kontrollbedingung zugeteilt. Je eine Treatment- und eine Kontrollklasse wurden von einer Lehrkraft unterrichtet. Insgesamt erklärten sich zehn Lehrer aus zehn verschiedenen Schulen (darunter sieben Männer und drei Frauen) dazu bereit, die Studie in Ihren Parallelklassen umzusetzen (vgl. Abbildung 51).

In Mehrebenenmodellen können auf jeder Ebene Regressionsgleichungen formuliert werden. Das Hauptziel dieser Arbeit besteht darin zu analysieren, ob Effekte auf das Treatment zurückgeführt werden können.

Im speziellen Fall dieser Studie lassen sich drei Ebenen unterscheiden: die Ebene der Messzeitpunkte, die Ebene der Individuen und die Ebene der Klassen.

Für die Untersuchung der Schulebene ist die Stichprobe mit zehn Aggregateinheiten zu klein.

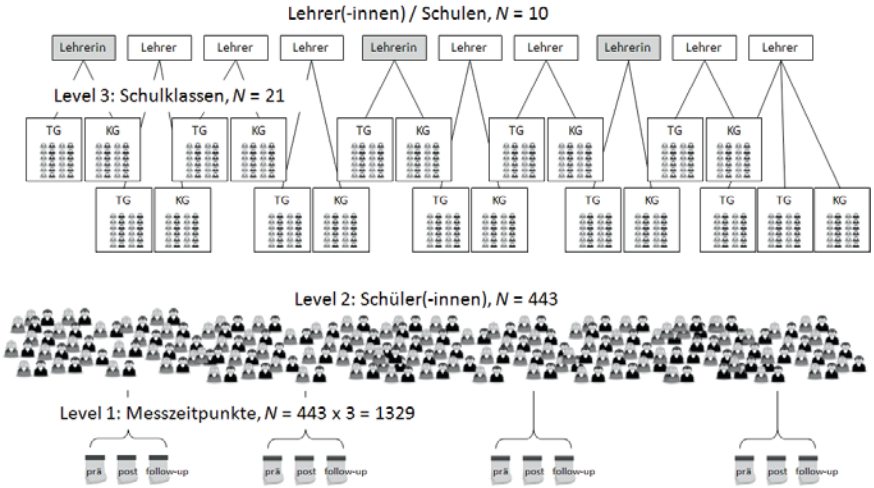


Abbildung 51: Veranschaulichung der Mehrebenenstruktur der Stichprobe, die Anzahl (*N*) auf Ebene 1 und 2 kann je nach Auswahl an Kovariaten und Messzeitpunkt variieren.

Das Hauptziel der Modellbildung bestand darin, zu analysieren, ob Effekte auf das Treatment zurückgeführt werden können. Hierzu wurde zunächst für jede abhängige Variable ein Mehrebenenmodell aufgestellt: Ein Modell zur Untersuchung des Effekts des Treatments auf die Leistung, ein Modell zur Untersuchung des Effekts des Treatments auf das konzeptuelle Verständnis und ein Modell zur Analyse der Wirkung auf die Motivation. Weitere Modelle ergaben sich aus vertiefenden und weiterführenden Fragestellungen, so der Frage nach der Wirkung des variierten Treatments und der Analyse möglicher ATI-Effekte (Aptitude-Treatment-Interaktionen, zur Erklärung des Begriffs sei auf Kapitel 2.3.9.8 Vertiefende Analysen zu Aptitude-Treatment-Interaktionen hingewiesen). Die Berechnung der Modelle erfolgte mit der freien Statistiksoftware „R“ unter Verwendung des Pakets „nlme“ [34], das von Pinheiro & Bates (2000) entwickelt wurde.

34 Letztes Update des Pakets Januar 2013 (vgl. Pinheiro & Bates, 2013).

In den folgenden Modellen wird die jeweilige interessierende Variable (Physikleistung, konzeptuelles Verständnis oder Schülermotivation) zum zweiten und dritten Messzeitpunkt in Abhängigkeit von den kognitiven Fähigkeiten, den vorherigen Leistungen (Noten), vom Geschlecht, vom Schultyp, von dem wahrgenommenen Lehrerengagements aus Sicht des jeweiligen Schülers und der Zuordnung zur Untersuchungsbedingung (Treatment- versus Kontrollgruppe) beschrieben.

Ziel des Modells ist es, die Änderung der jeweiligen interessierenden abhängigen Variablen in Relation zum ersten Messzeitpunkt unter Berücksichtigung relevanter Kovariaten zu beschreiben und auf statistische Signifikanz zu testen sowie die Stärke des Effekts zu schätzen.

Im Gegensatz zu klassischen Regressionsanalysen hat sich für Mehrebenenanalysen noch kein Standardverfahren für die Berechnung und den Bericht von Effektstärken etabliert (vgl. auch Hochweber, 2010, S. 153). Im Mittelpunkt dieser Arbeit steht die Analyse der Wirkung des Treatments auf die abhängigen Variablen Physikleistung bei repräsentationsbezogenen Aufgaben und auf das konzeptuelle Verständnis in Strahlenoptik. Zur Berechnung der Effektstärke wurde das Effektstärkemaß berechnet (vgl. Tymms, Merrel & Henderson, 1997; Tymms, 2004). Nach Tymms (2004, S. 56 f.) ergibt sich für dichotome Variablen aus der Differenz der betrachteten Mittelwerte (hier mittlerer Lernzuwachs) geteilt durch die gepoolte Standardabweichung (Wurzel aus der Varianz innerhalb der Gruppen). Angewandt auf die vorliegende Datenstruktur errechnet sich wie folgt:

$$\Delta_{\text{prä - post}} = \frac{\text{Differenz der Klassenmittelwerte des Zuwachses zwischen Treatment und Kontrollgruppe zu } t_2}{\sqrt{\text{Varianz innerhalb einer Klasse zu } t_2}}$$

$$= \frac{\beta_4}{\sqrt{\sigma^2_{\text{innerhalb Klasse zu } t_2}}}$$

$$= \frac{\beta_4}{\sqrt{\sigma_p^2 + g_2^2 \sigma_\varepsilon^2}}$$

Die Berechnung für die Effekte prä – follow-up erfolgte in entsprechender Weise. Analog zu Cohen's *d* gilt = 0.20 als kleiner, = 0.50 als mittlerer und = 0.80 als großer Effekt (vgl. Bortz & Döring, 2005, S. 568). Zudem wurde je Level (Messzeitpunkt, Individuum und Schulklasse) der Anteil erklärter Varianz berichtet, welcher sich auf das Treatment zurückführen lässt.

Des Weiteren wurde zum Vergleich der Wirkung von Variablen, bei denen sich die Bandbreite möglicher Werte unterscheidet (z.B. Vergleich der Wirkung

von Noten und Intelligenz auf die jeweilige abhängige Variable), sowohl die Kovariaten als auch die abhängigen Variablen standardisiert. Vollstandardisierte Koeffizienten (standardisiert bezüglich Prädiktor und Kriterium) werden in den folgenden Modellen mit „β“ angegeben, unstandardisierte Koeffizienten mit „b“. Liegt das Kriterium als dichotome Variable vor (wie etwa bei den Variablen *Bedingung* oder *Geschlechtszugehörigkeit*), werden ebenfalls bezüglich des Prädiktors standardisierte Koeffizienten mit „β“ angegeben (vgl. auch Fox, 1997, S. 153).

Da in den folgenden Modellen der Interzept die erwartete Leistung bzw. das konzeptuelle Verständnis eines Schülers darstellt, dessen Kovariaten den Wert 0 aufweisen, kann der Interzept für Variablen, die den Wert 0 nicht annehmen können, nicht sinnvoll interpretiert werden. Aus diesem Grund wurden alle metrischen Kovariaten grand-mean zentriert d.h. von allen verfügbaren Werten wurde der Mittelwert aller verfügbaren Werte abgezogen (vgl. Eid et al., 2011, S. 724; Raudenbush & Bryk, 2002, S. 31 ff.).

Nach der Zentrierung kann der Interzept als erwarteter Wert der abhängigen Variable eines absolut durchschnittlichen Schülers interpretiert werden, der sich in der Gruppe befindet, in der alle Dummy-kodierten kategorialen Variablen den Wert 0 annehmen (hier: Geschlecht = weiblich, Schultyp = IGS, Bedingung = Kontrollgruppe). Ein weiterer Vorteil der Zentrierung besteht darin, dass in Mehrebenenmodellen die Varianz des Interzepts interpretierbar wird (vgl. Hox, 2010, S. 62).

Im Gegensatz zu herkömmlichen Verfahren der linearen Regressionsanalyse ändert eine lineare Transformation der Kovariaten das Modell nicht essentiell (vgl. Eid et al, S. 724): Wird der Wert der Kovariate verdoppelt, halbiert sich der entsprechende Regressionskoeffizient. In Mehrebenenmodellen führt die Aufnahme von Random Slopes in Kombination mit linearen Transformationen (wie Grand-Mean-Centering) zu substantiellen Änderungen des Modells, die sich auf die inferenzstatistische Absicherung auswirken. Eine grafische Veranschaulichung dieses Problems findet sich in Hox (2010, S. 61). Da die im Folgenden aufgestellten Modelle keine Random-Slope-Terme enthalten, stellt sich das Problem in den vorgestellten Analysen nicht.

Mit fehlenden Werten wurde in den berechneten Modellen wie folgt umgegangen:

1. Umgang mit fehlenden Werten auf Ebene der Individuen und Messzeitpunkte:
   - Ein Vorteil des hier verwendeten Modells besteht darin, dass Werte von Schülern, die zu einem Messzeitpunkt fehlen, zu den anderen Messzeitpunkten berücksichtigt werden können, sofern alle relevanten Kovariaten

vorliegen. Wenn in späteren Analysen also Daten von $N = 443$ Schülern eingehen, bedeutet dies, dass pro Schüler alle berücksichtigten Kovariaten vorliegen und zu einem der drei Messzeitpunkte mindestens ein Test- bzw. Fragebogengesamtwert. Pro Messzeitpunkt betrachtet kann die Stichprobengröße von $N = 443$ abweichen.

- Zur Veranschaulichung ein Beispiel: liegen von einem Individuum Daten zu den kognitiven Fähigkeiten und den Schulnoten vor, jedoch keine Daten zu den Prätests, können die Daten der Posttests und Follow-up Tests (Physikleistung und konzeptuelles Verständnis) dennoch als abhängige Variablen in den Modellen für den zweiten und dritten Messzeitpunkt eingehen.

2. Umgang mit fehlenden Werten auf Itemebene:
   - Da es sich beim Physikleistungs-, beim Konzepttest und bei den Skalen des I-S-T 2000 R um Wissens- bzw. Fähigkeitstests handelt, wurden unbeantwortete Items mit 0 Punkten (nicht korrekt gelöst) bewertet.
   - Anders verhält es sich bei dem Motivationsfragebogen inklusive der Skala wahrgenommenes Lehrerengagement aus Schülersicht (LES): Sofern mindestens ein Item unbeantwortet blieb, wurden je Skala die jeweilige Antwort mittels einer multiplen linearen Regression vorhergesagt, unter der Voraussetzung, dass 2/3 der Items pro Skala beantwortet wurden: deterministic multiple regression imputation (vgl. Buck, 1960; Little & Rubin, 2002).
   - Im Fall der Skala LES (4 Items) wurde die jeweilige fehlende Antwort imputiert, sobald drei von vier Items beantwortet waren. Für die Skala „SK“ (Selbstkonzept) war diese Bedingung erfüllt, sobald sechs von neun Antworten vorlagen usw.
   - Die Imputation mittels multipler Regression stellt unter der Annahme von „MAR“ (missing at random)[35] einen optimalen Schätzer dar und ist der Mittelwertersetzung überlegen. Sind die Bedingungen für MAR nicht erfüllt, führt diese Methode zur Unterschätzung der Varianz des Kriteriums und zur Überschätzung der Kovarianzen zwischen Kriterium und Prädiktoren (vgl. Lüdtke & Robitzsch, 2010, S. 13).
   - Da kein Anlass bestand, der für eine anderweitige Annahme sprach, wurde davon ausgegangen, dass MAR vorliegt – zumal der Aufwand für komplexere Verfahren wie etwa die Multiple Imputation mittels des R-Pakets MICE (vgl.

35 Die Annahme für MAR ist gegeben, wenn nach Kontrolle der beobachteten Variablen, z.B. LES 1-LES 3 das Auftreten der fehlenden Variable, hier z.B. LES 4 nicht mehr von der (Gesamt-)Ausprägung des wahrgenommenen Lehrerengagement aus Schülersicht abhängt – also die Beobachtung zufällig fehlt (vgl. ebd., S. 13).

van Buuren & Groothuis-Oudshoorn, 2011) für Variablen, deren Untersuchung nicht im Mittelpunkt dieser Arbeit stehen, ungerechtfertigt erscheint.

- Unter Anwendung der Imputation mittels multipler Regression konnten statt $N = 387$ nun $N = 489$ Fälle im Prätest, statt $N = 362$ nunmehr $N = 477$ Fälle im Posttest und statt $N = 391$, $N = 495$ Fälle im Follow-up Test berücksichtigt werden.

2.3.9.2 Methodisches Vorgehen: Modellentwicklung

Zum Verständnis des Modells, welches dieses Ziel umsetzt, wird im Folgenden stufenweise von einem „leeren" Modell zu dem letztendlich verwendeten Modell übergegangen. Die Notation der jeweiligen Modelle erfolgt in Anlehnung an Fahrmeir, Kneib und Lang (2009, S. 254 ff.).

1) Modell 1 LM – Leeres Modell

$$y_{ij} = \beta_0 + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \varepsilon_{ij}$$

y_{ij}	Leistung des Individuums i zum Zeitpunkt j
β_o	Mittelwert der Leistung zum Zeitpunkt 1
i	Zeitpunkt (i = 1, 2, 3)
β_i	Beschreibt die Änderung der Leistung zwischen Zeitpunkt 1 und Zeitpunkt 2 bzw. Zeitpunkt 1 und Zeitpunkt 3. Dieses Vorgehen entspricht einem einfachen Einzelvergleich (einfacher Kontrast in SPSS) bei einer Varianzanalyse mit Messwiederholung (vgl. Bortz, 2005b, S. 306; Rudolf & Müller, 2004, S. 100)
t_2 / t_3	Dummy-Kodierung der Zeitpunkte $t_2 = \begin{cases} 1 & falls\ i = 2 \\ 0 & sonst \end{cases}$ $t_3 = \begin{cases} 1 & falls\ i = 3 \\ 0 & sonst \end{cases}$
j	Individuum (j = 1, ..., 525)
ε_{ij}	Residuum (Abweichung zwischen vorhergesagter Leistung und beobachteter Leistung des Individuums i zum Zeitpunkt j). Die ε_{ij} werden als identisch und unabhängig normalverteilt angenommen mit Mittelwert 0 und zu schätzender Standardabweichung σ.

Dieses Modell beschreibt nur die Leistung zu den unterschiedlichen Zeitpunkten ohne Berücksichtigung von Gruppenstrukturen, Kovariaten etc.
Um die Vergleichbarkeit mit den nachfolgenden gemischten Modellen zu gewährleisten, wurden die Modellparamter mittels der Maximum Likelihood Methode ermittelt.
Die Nullhypothese entspricht dabei der Annahme: „Es gab im Mittel keine Änderung der Leistung zwischen dem ersten und zweiten Messzeitpunkt". „H_0: $\beta_1 = 0$". Diese Hypothese kann im nlme-Paket mittels eines Wald-Tests geprüft werden.

Tabelle 43 Modell 1 LM – Leeres Modell

Variable		*b*	*SE*	$F_{(numDF,denDF)}$
β_0	Interzept	2.42	0.29	$4436.28_{(1, 1469)}^{***}$
β_1	Zuwachs prä – post	14.83	0.41	$689.15_{(1, 1469)}^{***}$
β_2	Zuwachs prä – follow-up	11.09	0.40	$752.71_{(1, 1469)}^{***}$

$^{***}p < .001$, $\sigma_\varepsilon = 6.34$

Aufbauend auf dem „leeren Modell" (vgl. Tabelle 43) wird nun ein Modell spezifiziert, welches berücksichtigt, dass die einzelnen Messwerte je einem Individuum zugeordnet sind. Dieses Modell entspricht einer Varianzanalyse mit Messwiederholung ohne Berücksichtigung von Kovariaten (vgl. Eid et al., 2011, S. 449 ff.).

2) Modell 2PM (Personenparameter) – Varianzanalyse mit Messwiederholung
$y_{ij} = \beta_0 + p_j + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \varepsilon_{ij}$

p_j Individuenparameter (geschätzt pro Individuum).

p_j wird als normalverteilt mit Mittelwert 0, zu schätzender Varianz und unabhängig von ε_{ij} angenommen (vgl. Hedeker, 2012, S. 3).

p_j ist der Unterschied in der Leistung der Person j über die Messzeitpunkte hinweg vom Mittelwert aller Leistungen aller Personen über die Messzeitpunkte hinweg (vgl. Eid et al., 2011, S. 450).

In diesem Modell wird die Korrelation der Messzeitpunkte im Individuum durch p_j adressiert. Die Varianz von p_j entspricht der Varianz zwischen den Individuen (vgl. Tabelle 44).

Tabelle 44 Modell 2 PM (Personenparameter) – Varianzanalyse mit Messwiederholung

Variable		*b*	*SE*	$F_{(numDF, denDF)}$
β_0	Interzept	2.26	0.28	$2392.87_{(1, 946)}^{***}$
β_1	Zuwachs prä – post	14.90	0.31	$1207.91_{(1, 946)}^{***}$
β_2	Zuwachs prä – follow-up	11.16	0.31	$1328.43_{(1, 946)}^{***}$

$^{***}p < .001$, $\sigma_p = 4.21$, $\sigma_\varepsilon = 4.76$

Die Intraklassen-Korrelation (*ICC*), abgekürzt $\hat{\rho}$, ergibt sich auf Ebene der Population wie folgt (vgl. Eid et al., 2011, S. 704):

$$\hat{\rho} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\varepsilon^2} = \frac{17.73}{17.73 + 22.67} = 0.44$$

p entspricht hier dem Individuenparameter und damit $\hat{\sigma}_p^2$ der Varianz zwischen den Individuen. Damit ist $\hat{\sigma}^2_{\text{Level-2}} = 17.73$.

ε_{ij} entspricht dem der Abweichung zwischen vorhergesagter Leistung und beobachteter Leistung des Individuums i zum Zeitpunkt *j*. Damit ist $\hat{\sigma}^2_{\text{Level-1}} = 22.67$.

In diesem Modell wurde noch nicht berücksichtigt, dass die Individuen je einer Schulklasse zugeordnet sind. Im folgenden Schritt soll geprüft werden, ob ein Modell, welches einbezieht, dass die Leistung von Individuen einer Klasse korreliert sein könnte, dem vorigen Modell überlegen ist.

3) Modell 3KM (Klassenparameter) – Varianzanalyse unter Berücksichtigung der Klassenstruktur

$y_{ijk} = \beta_0 + u_k + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \varepsilon_{ij}$

u_k	Klassenparameter (geschätzt pro Schulklasse)
	u_k wird als normalverteilt mit Mittelwert 0 zu schätzender Varianz und unabhängig von ε_{ij} angenommen.

In Anlehnung an Modell 2 PM (Personenparameter) – Varianzanalyse mit Messwiederholung (siehe Tabelle 44) wird nun die Korrelation der Individuen inner-

halb der Schulklasse untersucht.[36] Das Modell berücksichtigt mit dem Random Interzept u_k, dass die in der Stichprobe enthaltenen Schulklassen lediglich eine Auswahl an Schulklassen aus der Grundgesamtheit darstellt (vgl. Tabelle 45). Durch u_k erhalten wir einen Term, der eine Korrelationsstruktur der Individuen (*ICC*), die jeweils einer Schulklasse angehören, schätzt. In dem aufgestellten Modell entspricht die Korrelation der Leistungen der Individuen einer Schulklasse untereinander nun gerade dem *ICC* (vgl. Hochweber, 2010, S. 144). Die Varianz von u_k ist gleich der Varianz zwischen den Schulklassen.

Tabelle 45 Modell 3 KM (Klassenparameter) – Varianzanalyse unter Berücksichtigung der Klassenstruktur

Variable		*b*	*SE*	$F_{(numDF,\ denDF)}$
β_0	Interzept	2.12	0.75	$224.42_{(1,\ 1449)}^{***}$
β_1	Zuwachs prä – post	15.02	0.35	$942.21_{(1,\ 1449)}^{***}$
β_2	Zuwachs prä – follow-up	11.22	0.35	$1026.50_{(1,\ 1449)}^{***}$

$^{***}p < .001$, $\sigma_u = 3.25$, $\sigma_\varepsilon = 5.50$

Die Intraklassen-Korrelation ergibt sich wie folgt (vgl. Eid et al., 2011, S. 704)

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2} = \frac{10.55}{10.55 + 30.23} = 0.26$$

u_k entspricht hier dem Klassenparameter. Damit ist $\hat{\sigma}^2_{\text{Level-2}} = 10.55$.

ε_{ij} entspricht wieder dem der Abweichung zwischen vorhergesagter Leistung und beobachteter Leistung des Individuums i zum Zeitpunkt j. Damit ist $\hat{\sigma}^2_{\text{Level-1}} = 30.23$.

Im Ergebnis zeigt sich, dass der *ICC* dabei hinreichend hoch ist, um die Berücksichtigung der Klassenstruktur durch einen Random Interzept in der Schulklasse zu rechtfertigen (vgl. Eid et al., 2011, S. 705). Werden beide Ebenen (Messzeitpunkte und Schulklassen) berücksichtigt, ergibt sich folgendes Modell:

36 Dieses Modell berücksichtigt nicht die Individuenstruktur, es werden lediglich die Messungen zu den verschiedenen Zeitpunkten miteinander verglichen (ohne den Messungen Individuen zuzuordnen).

4) Modell 4 KPM (Klassen- und Personenparameter) – Varianzanalyse unter Berücksichtigung von sowohl der Klassenstruktur als auch der Individuenkomponente

$$y_{ik} = \beta_o + p_j + u_k + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \varepsilon_{ij}$$

u_k / p_j	Klassenparameter (geschätzt pro Schulklasse) / Personenparameter (geschätzt pro Individuum)
	Beide Parameter werden als normalverteilt mit Mittelwert 0, zu schätzender Varianz und unabhängig voneinander und ε_{ij} angenommen.

Verschiedene Autoren wie Hox und Kreft (1994), Hox (2002) sowie Maas und Hox (2004) fordern überwiegend eine Fallzahl von $N \geq 50$ auf höheren Ebenen. Allerdings bezieht sich diese Forderung insbesondere auf Fragestellungen, in denen explizit die Interaktion zwischen den Ebenen untersucht werden soll. Aufgrund der besseren Schätzung der Messfehler und der Möglichkeit, durch die Mehrebenenstruktur der Verletzung der Unabhängigkeitsannahme gerecht zu werden, wurde die Klassenstruktur in das Modell aufgenommen (vgl. Tabelle 46). Die Ergebnisse sind somit mit dem hier untersuchten Datensatz von $N = 21$ (Schulklassen) nur unter dem Vorbehalt zu interpretieren, dass die Schätzung des Einflusses der Schulklasse mit Unsicherheit behaftet ist.

Folgende Gründe sprachen dafür, diesen Vorbehalt in Kauf zu nehmen:

- Erstens geht es in der Modellbildung vorwiegend um den Einfluss des Treatments (UV = Bedingung). Somit soll der Einfluss der Schulklasse kontrolliert werden und ist nicht selbst Gegenstand der Forschungsfrage. Gemäß Eid et al. (2011, S. 715) genügt bereits ein Stichprobenumfang von $N = 10$ Einheiten der Aggregatebene, sofern „man lediglich an einer Schätzung der Parameter im festen Teil des Modells interessiert (ist)“ (zit.n. ebd.). Daher wird angenommen, dass die fixen Effekte (wie der Einfluss des Treatments) zuverlässig geschätzt werden.
- Zweitens ergibt sich bereits für einen *ICC* von .10 bei $N = 10$ auf der Aggregatebene ein 50 % Risiko einer statistischen Fehlentscheidung (vgl. Eid et al., 2011, S. 705).

Nach Einschätzung der Autorin wiegt der Vorteil somit die Nachteile auf. Zu den Ergebnissen sei auf Tabelle 46 hingewiesen.

Tabelle 46 Modell 4KPM (Klassen- und Personenparameter) – Varianzanalyse unter Berücksichtigung von sowohl der Klassenstruktur als auch der Individuenkomponente

Variable		*b*	*SE*	$F_{(numDF, denDF)}$
β_0	Interzept	2.09	0.75	$222.33_{(1, 946)}^{***}$
β_1	Zuwachs prä – post	14.98	0.31	$1236.87_{(1, 946)}^{***}$
β_2	Zuwachs prä – follow-up	11.21	0.30	$1355.58_{(1, 946)}^{***}$

*** $p < .001$, $\sigma_p = 2{,}81$, $\sigma_u = 3.22$, $\sigma_\varepsilon = 4.74$

Die *ICCs* auf den entsprechenden Ebenen erscheinen hinreichend groß, um sowohl die Schachtelung der Messzeitpunkte als auch die Schachtelung der Individuen in Klassen zu rechtfertigen:

- Die Varianz innerhalb des Individuums entspricht der Varianz der Residuen $\hat{\sigma}_\varepsilon^2 = 22.56$
- Die Varianz zwischen den Individuen einer Klasse ist gleich der Varianz des Random Interzept auf Ebene des Individuums $\hat{\sigma}_p^2 = 7.87.$
- Die Varianz zwischen den Klassen entspricht der Varianz des Random Interzept auf Ebene der Schulklasse $\hat{\sigma}_u^2 = 10.55$

Die Intraklassen-Korrelation ergibt sich auf Ebene der Population wie folgt (vgl. Eid et al., 2011, S. 704):

ICC Level 3: $$\widehat{\rho_3} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_p^2 + \hat{\sigma}_\varepsilon^2} = \frac{10.55}{10.55 + 7.87 + 22.56} = 0.25$$

= „Ähnlichkeit der Individuen einer Klasse"

ICC Level 2: $$\widehat{\rho_2} = \frac{\hat{\sigma}_u^2 + \hat{\sigma}_p^2}{\hat{\sigma}_u^2 + \hat{\sigma}_p^2 + \hat{\sigma}_\varepsilon^2} = \frac{10.55 + 7.87}{10.55 + 7.87 + 22.56} = 0.45$$

= „Ähnlichkeit der Messzeitpunkte eines Individuums"

5) Modell 5 KPM&H – Varianzanalyse unter Berücksichtigung von sowohl der Klassenstruktur als auch der Individuenkomponente unter Adressierung der Heteroskedastizität

Die Betrachtung der Boxplots der Ergebnisse des Leistungstests zu den verschiedenen Messzeitpunkten lässt unterschiedliche Varianzen (Heteroskedastizität) erkennen (siehe *Abbildung 53*, S. 289). So zeigt sich in den Ergebnisse des Prätests ein starker „Bodeneffekt" und damit verbunden eine geringere Varianz (siehe auch die nachfolgenden Histogramme, Abbildung 52):

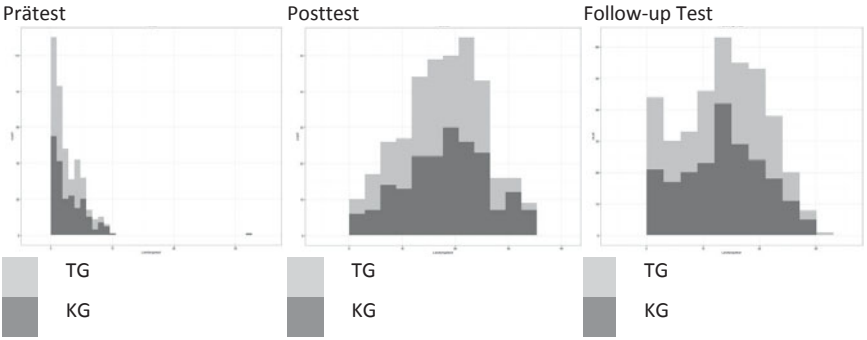


Abbildung 52: Histogramme der Leistungstestergebnisse je Messzeitpunkt, TG = Treatmentgruppe, KG = Kontrollgruppe

Der Grund besteht sehr wahrscheinlich darin, dass die Schüler zum ersten Messzeitpunkt nur geringe Vorkenntnisse mitbringen, da die Bildentstehung an der Sammellinse noch ein neues Thema ist. Diese geringen Vorkenntnisse konnten vermutlich auch nicht durch früheren Physikunterricht aufgefangen werden, weil es sich sowohl für die Siebt- als auch für die Achtklässler um das erste Schuljahr des Physikunterrichts handelt. Nach der Unterrichtsreihe fällt die Varianz entsprechend größer aus, wobei sich die Verteilungen im Post- und im Follow-up Test ähneln.

Aufgrund der Heteroskedastizität wird im nächsten Schritt zusätzlich die Möglichkeit unterschiedlicher Varianzen in den Fehlertermen je Messzeitpunkt aufgenommen, welche durch die Gewichtskomponente g geschätzt wird (siehe auch Tabelle 47).

Durch die Hinzunahme der Fehlergewichte können keine erheblichen Änderungen der geschätzten fixen Effekte des Modells beobachtet werden, wohl aber eine Verbesserung des Modells, wie aus Tabelle 49 ersichtlich wird, welche einen Überblick zu den Informationskriterien bietet.

$y_{ij} = \beta_o + p_j + u_k + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \varepsilon_{ij}$

wobei Varianz von ε_{ij} gleich $g_i^2 \cdot \sigma_\varepsilon^2$ ist.

g_i Gewichtskomponente. Die Gewichtskomponenten wurden mittels des Befehls „weights" geschätzt, der sich dafür eignet, Heteroskedastizität innerhalb von Gruppen abzubilden.[37]

$g_1 = 1$, g_2 und g_3 werden geschätzt.

g_2 beschreibt den (multiplikativen) Unterschied der Standardabweichung der Fehler zum Zeitpunkt 1 (σ_ε) im Vergleich zu Zeitpunkt 2, g_3 analog dazu den (multiplikativen) Unterschied der Standardabweichung der Fehler zum Zeitpunkt 1 im Vergleich zu Zeitpunkt 3.

Man beachte, dass im Programmpaket nlme nicht die Standardabweichungen der Fehlerterme je Messzeitpunkt ausgegeben werden, sondern die Standardabweichung zum Messzeitpunkt 1 und die Gewichte für die anderen Messzeitpunkte. Die Standardabweichung zum Messzeitpunkt 2 bzw. 3 ergibt sich als Produkt aus der Standardabweichung zum Messzeitpunkt 1 und dem Gewicht des Zeitpunktes 2 bzw. 3.

Tabelle 47 Modell 5 KPM&H – Varianzanalyse unter Berücksichtigung von sowohl der Klassenstruktur als auch der Individuenkomponente unter Adressierung der Heteroskedastizität

Variable		*b*	*SE*	$F_{(numDF, denDF)}$
β_0	Interzept	2.30	0.39	91.51 (1, 946)
β_1	Zuwachs prä – post	14.90	0.32	2001.67 (1, 946)
β_2	Zuwachs prä – follow-up	11.15	0.30	1336.99 (1, 946)

*** $p < .001$, $\sigma_p = 1{,}75$, $\sigma_u = 1{,}71$, $\sigma_\varepsilon = 1{,}59$, $g_1 = 1$, $g_2 = 4.26$, $g_3 = 4.12$

6) Modell 6 KPM&H&Korr – Varianzanalyse unter Berücksichtigung von sowohl der Klassenstruktur als auch der Individuenkomponente unter Adressierung der Heteroskedastizität mit unstrukturierter Korrelationsmatrix der Fehlerterme im Individuum

Die durch den Individuenparameter p_j alleine implizierte Korrelation der Leistungen zu den drei Zeitpunkten innerhalb eines Individuums erscheint zu starr. Das R Paket nlme erlaubt im Gegensatz zu dem R Paket lme4 die Schätzung

37 „Weights: an optional varFunc object or one-sided formula describing the within-group heteroscedasticity structure" (zit. n. Pinheiro & Bates, 2013, S. 145).

von beliebigen Korrelationen der Fehler im Individuum, welches wir in dem Modell mit aufnehmen (vgl. Tabelle 48 zu den Ergebnissen). Die Informationskriterien des *AIC* und *BIC* stützen dieses Vorgehen (siehe Tabelle 49).

$$y_{ij} = \beta_o + p_j + u_k + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \varepsilon_{ij}$$

$$= \beta_o + p_j + u_k + \varepsilon_{1j} \cdot t_1 + (\beta_1 + \varepsilon_{2j}) \cdot t_2 + (\beta_2 + \varepsilon_{3j}) \cdot t_3 \qquad \text{(identische Umformulierung)}$$

ε_{ij}

(Random Variable)

Fehlerterm: Abweichung der tatsächlich gemessenen Leistung des Schülers j der Klasse k zum Zeitpunkt i von der durch das Modell vorhergesagten Leistung.

ε_{ij} ist für jede Kombination aus i/j Realisation einer (univariat) normalverteilten Zufallsvariablen mit Mittelwert $\mu = 0$ und Standardabweichung $g_i \cdot \sigma_\varepsilon$. Die Standardabweichung ist damit für jeden Zeitpunkt i = 1, 2, 3 unterschiedlich.

(man beachte, dass $g_1 = 1$ ist / siehe oben)

Damit adressiert der Fehler über die Gewichte g_i die unterschiedlichen Varianzen der Leistungstestergebnisse zu den verschiedenen Messzeitpunkten (Heteroskedastizität).

Ferner wird die Korrelation der Messzeitpunkte untereinander berücksichtigt, welche sich aus dem repeated Measurement Design ergeben. Dies ist zwar bereits durch den Individuenparameter p adressiert. Durch diesen ist jedoch die Korrelation der Messzeitpunkte immer gleich, wohingegen die beliebige Korrelation der Fehlerterme höhere Flexibilität bietet.

Man beachte auch die Darstellung:

$$\varepsilon_{ij} = t_1 \, \varepsilon_{1j} + t_2 \, \varepsilon_{2j} + t_3 \, \varepsilon_{3j}$$

Für die Korrelation der Fehlerterme ergibt sich:

$$\begin{pmatrix} 1 & 0.15 & 0.18 \\ 0.15 & 1 & 0.61 \\ 0.18 & 0.61 & 1 \end{pmatrix}, \text{ d.h. } \begin{aligned} \rho_{12} &= 0.15 \\ \rho_{13} &= 0.18 \\ \rho_{23} &= 0.61 \end{aligned}$$

Die Korrelation der Leistungen eines Individuums zu den Messzeitpunkten ergibt sich im Modell nun als Kombination der Effekte durch die Individuenkomponente p_j und der Korrelation der Fehlerterme.

Als Element zur Erzeugung einer Korrelationsstruktur innerhalb des Individuums benötigt man die Individuenkomponente p_j nun im Grunde nicht

mehr. Die Komponente wurde dennoch beibehalten, um einerseits die Gruppenstruktur der Messungen weiterhin deutlich zu machen und andererseits die Varianz zwischen den Individuen weiterhin abschätzen zu können. Eine Herausnahme des Terms verbessert jeweils *AIC* und *BIC* bei gleichbleibender *Devianz* und Loglikelihood, wobei sich die Korrelationsschätzungen der Fehlerterme entsprechend anpassen. Die durch p_j „erklärte" Varianz wird dann durch höhere Fehlervarianzen aufgefangen.

Tabelle 48 Modell 6 KPM&H&Korr – Varianzanalyse unter Berücksichtigung von sowohl der Klassenstruktur als auch der Individuenkomponente unter Adressierung der Heteroskedastizität mit unstrukturierter Korrelationsmatrix der Fehlerterme im Individuum

Variable		*b*	*SD*	$F_{(numDF,\ denDF)}$
β_0	Interzept	2.33	0.37	$65.29_{(1,\ 946)}^{***}$
β_1	Zuwachs prä – post	14.75	0.32	$955.74_{(1,\ 946)}^{***}$
β_2	Zuwachs prä – follow-up	11.01	0.31	$1274.79_{(1,\ 946)}^{***}$

*** $p < .001$, $\sigma_p = 1{,}76$, $\sigma_u = 1{,}61$, $\sigma_\varepsilon = 1{,}97$, $g_1 = 1$, $g_2 = 3.62$, $g_3 = 3.53$

Zur Beurteilung der Modellgüte wurden die Informationskriterien *AIC* und *BIC* sowie die *Devianz* des Modells herangezogen. Beim *AIC* und *BIC* handelt es sich um Maße der Anpassungsgüte des Modells an die vorliegenden Daten unter Berücksichtigung der Modellkomplexität. *AIC* und *BIC* sind Maßzahlen für den Informationsverlust. Beide errechnen sich aus der logarithmierten Likelihood-Funktion und einem Strafterm, dessen Größe von der Anzahl der geschätzten Parameter abhängt. Im Gegensatz zum *AIC* fallen beim *BIC* komplexere Annahmen stärker ins Gewicht. Weniger sparsame Modelle werden beim *BIC* also stärker bestraft als beim *AIC* (vgl. Moosbrugger & Kelava, 2007, S. 390).
AIC und *BIC* lassen sich nicht direkt interpretieren, sondern können nur als relative Werte innerhalb eines Modellvergleichs auf Basis derselben Stichprobe zueinander in Beziehung gesetzt werden. Die *Devianz* stellt ein weiteres Kriterium dar mit dessen Hilfe die Passung des Modells auf die jeweiligen Daten beurteilt werden kann. Sie wird aus der maximierten doppelten negativen Log-Likelihood berechnet (vgl. Hox, 2002, S. 42; Eid et al., 2011, S. 715):

$$DEV = -2\ln(L)$$

Auch für die *Devianz* gilt, je größer der Wert, desto schlechter ist die Passung des Modells. Wie aus der Tabelle 49 ersichtlich ist, zeichnet sich das letzte Modell (KPM&H&Korr) durch die besten Modelleigenschaften aus: So weist es bezüglich des *AIC* (Akaike's Informationskriterium), des *BIC* (Bayesian Informationskriterium) und der *Devianz* die geringsten Werte auf. Daher wurde das Modell, welches sowohl die Klassenstruktur, die Messwiederholung durch die Individuenkomponente als auch die Heteroskedastizität berücksichtigt und eine unstrukturierte Korrelationsmatrix der Fehlerterme im Individuum zugrunde legt, als Ausgangsmodell gewählt.

Tabelle 49 Überblick über die Modelleigenschaften bei steigender Komplexität

Nr.	Modell-kürzel	Berücksichtigte Parameter	*AIC*	*BIC*	*Devianz*	σ^2_u	σ^2_p	σ^2_ε *zu* t_1	σ^2_ε *zu* t_2	σ^2_ε *zu* t_3
1	LM	Leeres Modell	9623.66	9644.83	9615.66	-	-		40.20	
2	PM	LM + Personen-parameter (PM)	9386.67	9413.14	9376.67	-	17.73		22.67	
3	KM	LM + Klassen-parameter (KM)	9273.04	9299.52	9263.04	10.55	-		30.23	
4	KPM	LM + KM + PM	9192.13	9223.90	9180.13	10.35	7.87		22.56	
5	KPM&H	KPM + Heteros-kedastizität (H)	8850.14	8892.49	8834.13	2.92	3.05	2.54	46.22	43.20
6	KPM&H &Korr	KPM&H + unstruktu-rierte Korrelation der Fehler zu den Mess-zeitpunkten	8669.04	8727.28	8647.04	2.60	1.38	3.88	51.01	48.45

Tabelle 50 Überblick zum Einfluss der Kovariaten auf Wissen und Problemlösen bei repräsentationsbezogenen Aufgaben (Physikleistung)

	Prätest			Posttest			Follow-up Test		
	b	*β*	$F_{(numDF,\ denDF)}$	*b*	*β*	$F_{(numDF,\ denDF)}$	*b*	*β*	$F_{(numDF,denDF)}$
Notenfaktor PCA (D, M, Ph)	-0.39	-0.06	30.35 $_{(1,\ 496)}$***	-2.30	-0.36	122.45 $_{(1,\ 936)}$***	-2.21	-0.34	119.18 $_{(1,\ 936)}$***
IQ-verbal (Satzergänzung)	0.05	0.06	30.01 $_{(1,\ 458)}$***	0.17	0.20	30.39 $_{(1,\ 873)}$***	0.13	0.16	20.47 $_{(1,\ 873)}$***
IQ-räumlich (Würfelaufgaben)	0.00	0.00	0.05 $_{(1,\ 458)}$	0.18	0.18	25.81 $_{(1,\ 873)}$***	0.13	0.13	13.10 $_{(1,\ 873)}$**
IQ-figural-logisch (Matrizen)	0.02	0.02	3.03 $_{(1,\ 456)}$	0.15	0.16	18.80 $_{(1,\ 869)}$***	0.13	0.14	15.16 $_{(1,\ 869)}$**
LES[a]	0.00	0.00	0.02 $_{(1,\ 465)}$	0.17	0.09	6.38 $_{(1,\ 900)}$*	0.02	0.01	0.11 $_{(1,\ 900)}$
Geschlecht	0.04	0.00	0.06 $_{(1,\ 502)}$	0.49	0.06	0.60 $_{(1,\ 942)}$	-0.91	-0.10	2.22 $_{(1,\ 942)}$
Klassengröße	0.02	0.01	0.05 $_{(1,\ 18)}$	0.34	0.11	1.61 $_{(1,\ 942)}$	0.38	0.13	14.80 $_{(1,\ 942)}$**
Schultyp	1.88	0.21	7.24 $_{(1,\ 18)}$*	7.86	0.88	104.27 $_{(1,\ 942)}$***	6.27	0.70	66.04 $_{(1,\ 942)}$***

* $p < .05$, ** $p < .01$, *** $p < .001$

[a]Lehrerengagement aus Schülersicht

Nachdem das Ausgangsmodell festgelegt ist, steht die Überlegung an, welche Kovariaten in das Modell eingehen sollen. Basierend auf dem Ausgangsmodell wurde jeweils eine Kovariate hinzugenommen und analysiert, ob die jeweilige Kovariate je Messzeitpunkt einen signifikanten Erklärungswert für die abhängige Variable aufweist. Die betrachteten Modelle sehen damit jeweils wie folgt aus:

$$y_{ij} = \beta_o + p_j + u_k + \beta_1 \cdot t_2 + \beta_2 \cdot t_3 + \beta_3 \cdot x + \beta_4 \cdot t_2 \cdot x + \beta_5 \cdot t_3 \cdot x + \varepsilon_{ij}$$

x jeweilige Kovariate

Die Analyse zeigt, dass sich insbesondere die kognitiven Fähigkeiten (Subskalen des I-S-T 2000 R), die vorherigen Schulleistungen (Noten) und der Schultyp signifikant auf das Abschneiden im Leistungstest zu allen drei Messzeitpunkten auswirken (vgl. Tabelle 50). Das Lehrerengagement aus Sicht der Schüler hat einen signifikanten Einfluss auf den Lernzuwachs prä – post ($b = 0.17$, $\beta = 0.09$, $p < .01$). Die Klassengröße wirkt sich lediglich signifikant auf den Lernzuwachs prä – follow-up aus ($b = 0.38$, $\beta = 0.13$, $p < .01$). Da somit jede der infrage kommenden Variablen außer der Geschlechtszugehörigkeit zu einem der drei Zeitpunkte einen signifikanten Erklärungswert aufweist, wurden alle Variablen bis auf das Geschlecht in das Modell aufgenommen.

Unter Berücksichtigung der genannten Variablen ergibt sich nun folgendes Gesamtmodell:

$$\begin{aligned} y_{ijk} = {} & \beta_0 + p_j + u_k + \beta_1 t_2 + \beta_2 t_3 \\ & + \beta_3 \text{Bedingung}_k + \beta_4 t_2 \text{Bedingung}_k + \beta_5 t_3 \text{Bedingung}_k \\ & + \beta_6 \text{Notenfaktor}_j + \beta_7 t_2 \text{Notenfaktor}_j + \beta_8 t_3 \text{Notenfaktor}_j \\ & + \beta_9 \text{IQ}_{\text{verbal j}} + \beta_{10} t_2 \text{IQ}_{\text{verbal j}} + \beta_{11} t_3 \text{IQ}_{\text{verbal j}} \\ & + \beta_{12} \text{IQ}_{\text{figural-räumlich j}} + \beta_{13} t_2 \text{IQ}_{\text{figural-räumlich j}} + \beta_{14} t_3 \text{IQ}_{\text{figural-räumlich j}} \\ & + \beta_{15} \text{IQ}_{\text{figural-logisch j}} + \beta_{16} t_2 \text{IQ}_{\text{figural-logisch j}} + \beta_{17} t_3 \text{IQ}_{\text{figural-logisch j}} \\ & + \beta_{18} \text{Klassengröße}_k + \beta_{19} t_2 \text{Klassengröße}_k + \beta_{20} t_3 \text{Klassengröße}_k \\ & + \beta_{21} \text{Schultyp}_k + \beta_{22} t_2 \text{Schultyp}_k + \beta_{23} t_3 \text{Schultyp}_k \\ & + \beta_{24} \text{LES}_j + \beta_{25} t_2 \text{LES}_j + \beta_{26} t_3 \text{LES}_j \\ & + \varepsilon_{ij} \end{aligned}$$

Erklärungen zu den Variablen:

i	Zeitpunkt (i = 1,2,3).
j	Individuum (j = 1,...,525).

k	Klasse (k = 1,...,21).
y_{ijk} (abhängige Variable)	Gemessene Leistung des Individuums j aus der Klasse k zum Zeitpunkt i.
β_o (fixed variable)	Geschätzte Leistung einer durchschnittlichen weiblichen Schülerin, aus der Kontrollbedingung, welche eine durchschnittlich große Klasse einer Gesamtschule besucht und auch bezüglich kognitiver Fähigkeiten und Noten durchschnittlich abschneidet: der Mittelwert der Kovariaten Lehrerengagement aus Schülersicht (LES), IQ_{verbal}, $IQ_{figural-räumlich}$, $IQ_{figural-logisch}$, Notenfaktor beträgt jeweils 0, da die genannten Variablen grand-mean-zentriert wurden.
p_j (Random Variable)	Mittlerer Unterschied der Leistungen des Schülers j zum Schnitt seiner Klasse. Dieser mittlere Unterschied wird über alle Messzeitpunkte hinweg geschätzt. D.h. wenn für j = 1 die geschätzte Größe $p_1 = 2$ ist, so gehen wir davon aus, dass der Schüler zu allen Messzeitpunkten um 2 Punkte besser ist, als seine Klassenkameraden (bei gleichen Werten der Kovariaten). p_j ist für jeden Schüler j eine Realisation einer normalverteilten Zufallsvariablen mit Mittelwert $\mu = 0$ und Standardabweichung $\sigma = \sigma_p$. p_j ist unabhängig von allen anderen Random Termen des Modells. Die Komponente p_j ermöglicht zusammen mit der unstrukturierten Korrelationsmatrix der Fehlerterme ε (siehe unten) die Korrelation der Messzeitpunkte i_1 und i_2 innerhalb eines Individuums j. Man kann errechnen: $$\rho(y_{i_1\,j\,k}, y_{i_2\,j\,k}) = \frac{\sigma_p{}^2+\sigma_u{}^2+\rho_{i_1 i_2}\cdot g_{i_1}\cdot g_{i_2}\cdot\sigma_\varepsilon{}^2}{\sqrt{\sigma_u{}^2+\sigma_p{}^2+g_{i_1}^2\cdot\sigma_\epsilon{}^2\cdot}\sqrt{\sigma_u{}^2+\sigma_p{}^2+g_{i_2}^2\cdot\sigma_\epsilon{}^2\cdot}}$$ σ_u^2 Varianz der Zufallsvariablen u σ_p^2 Varianz der Zufallsvariablen v σ_ε^2 Varianz der Zufallsvariablen ε zum Zeitpunkt 1 g_{i1}/g_{i2} Gewicht des Zeitpunkts i_1/i_2 (siehe unten) p_j adressiert damit Unterschiede zwischen Schülern, welche auf nicht erhobene Kovariaten (z.B. sozio-ökonomischer Status der Eltern, soziale Kompetenz, sonstige unbekannte Einflüsse) zurückzuführen sind.

u_k (Random Variable)	Mittlerer Unterschied der Leistungen von Schülern in Klasse k zum Schnitt über alle Klassen. Dieser mittlere Unterschied wird über alle Messzeitpunkte hinweg geschätzt. D.h. wenn für k = 1 die geschätzte Größe $u_1 = 2$ ist, so gehen wir davon aus, dass alle Schüler der Klasse 1 zu allen Messzeitpunkten um 2 Punkte besser sind, als das Modell es andernfalls vorhersagen würde. u_k ist für jede Klasse k eine Realisation einer normalverteilten Zufallsvariablen mit Mittelwert $\mu = 0$ und Standardabweichung $\sigma = \sigma_u$; u_k ist unabhängig von allen anderen Random Termen des Modells; u_k sorgt für die Korrelation ϱ der Leistungen y aller Individuen $j_1, j_2, \ldots$ einer Klasse k zum selben Messzeitpunkt i: $$\rho(y_{i\,j_1\,k}, y_{i\,j_2\,k}) = \frac{{\sigma_u}^2}{{\sigma_u}^2 + {\sigma_p}^2 + {\sigma_{\varepsilon_i}}^2}$$ σ_u^2 Varianz der Zufallsvariablen u σ_p^2 Varianz der Zufallsvariablen v $\sigma_{\varepsilon i}^2$ Varianz der Zufallsvariablen ε zum Zeitpunkt i (= $g_i^2 \sigma_\varepsilon^2$) u_k adressiert damit Unterschiede zwischen Klassen, welche auf nicht erhobene Kovariaten (z.B. Klassenklima, regionale Einflüsse wie z.B wohnortsgebundener sozioökonomischer Status, Stadt vs. Land etc.) zurückzuführen sind.
t_1 (Dummy-Kodierung)	$= \begin{cases} 1 & falls\ i = 1 \\ 0 & sonst \end{cases}$
t_2 (Dummy-Kodierung)	$= \begin{cases} 1 & falls\ i = 2 \\ 0 & sonst \end{cases}$
t_3 (Dummy-Kodierung)	$= \begin{cases} 1 & falls\ i = 3 \\ 0 & sonst \end{cases}$
β_1 (fixed variable)	Mittlere Änderung der Leistung (= Leistungszuwachs) zwischen Zeitpunkt i = 1 und Zeitpunkt i = 2. Mittlere Änderung bedeutet dabei: die Änderung der Leistung, wenn von einer „durchschnittlichen“ Schülerin (siehe β_0) ausgegangen wird.

β_2 (fixed variable)	Mittlere Änderung der Leistung (= Leistungszuwachs) zwischen Zeitpunkt i = 1 und Zeitpunkt i = 3. Mittlere Änderung bedeutet dabei: die Änderung der Leistung, wenn von einer „durchschnittlichen" Schülerin (siehe β_0) ausgegangen wird.
Bedingung_k (Dummy Kodierung / unabhängige Variable)	$= \begin{cases} 1 & falls\ Klasse\ k\ in\ der\ Treatmentgruppe\ ist \\ 0 & sonst \end{cases}$
β_3 Bedingung_k (fixed variable)	„Einfluss" der Variable Bedingung zum Zeitpunkt i = 1. Da zu diesem Zeitpunkt noch kein Einfluss des Treatments vorliegt, misst β_{13} den Leistungsunterschied zwischen Treatment- und Kontrollgruppe zum Zeitpunkt i = 1.
β_4 Bedingung_k (fixed variable)	Einfluss der Variable Bedingung auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zum Zeitpunkt i = 1. Bsp.: Es sei $\beta_4 = 2$. Dann sagt das Modell für Schüler aus der Treatmentgruppe zum Zeitpunkt i = 2 einen um 2 Punkte *höheren Leistungszuwachs* zwischen i = 1 und i = 2 voraus, als in der Kontrollgruppe.
β_5 Bedingung_k (fixed variable)	Einfluss der Variable Bedingung auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zum Zeitpunkt i = 1. Bsp.: Es sei $\beta_5 = 2$. Dann sagt das Modell für Schüler aus der Treatmentgruppe zum Zeitpunkt i = 3 einen um 2 Punkte *höheren Leistungszuwachs* zwischen i = 1 und i = 3 voraus, als in der Kontrollgruppe.
Notenfaktor_j (Kovariate)	Gewichtete Summe aus der Deutsch-, Physik- und Mathematiknote des Schülers j. Entspricht der ersten Komponente einer Principal Component Analysis der drei Noten. Der aus den Einzelnoten berechnete Notenfaktor wurde grand-mean-zentriert.

β_6 Notenfaktor$_j$ (fixed variable)	Einfluss der Noten auf die Leistung.
β_7 Notenfaktor$_j$ (fixed variable)	Zusätzlicher Einfluss der Noten auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1.
β_8 Notenfaktor$_j$ (fixed variable)	Zusätzlicher Einfluss der Noten auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zu Zeitpunkt i = 1.
$IQ_{verbal\ j}$ (Kovariate)	(Grand-mean-)zentrierte verbale Intelligenz des Schülers j: Bsp.: Es sei 105 die gemessene verbale Intelligenz des Schülers j und der Durchschnitt der verbalen Intelligenz sei 99. Dann ist $IQ_{verbal\ j} = 6$
β_9 $IQ_{verbal\ j}$ (fixed variable)	Einfluss der verbalen Intelligenz auf die Leistung. Bsp.: Es sei $\beta_3 = 0.5$. Dann erhöht sich die Vorhersage der Leistung des Schülers zum Zeitpunkt i = 1 um 0.50 Punkte je zusätzlichem Punkt IQ_{verbal}
β_{10} $IQ_{verbal\ j}$ (fixed variable)	Zusätzlicher Einfluss der verbalen Intelligenz auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1. Bsp.: Es sei $\beta_4 = 0.50$. Dann erhöht sich die Vorhersage der Leistungszunahme des Schülers zum Zeitpunkt i = 2 um Vergleich zu Zeitpunkt i = 1 um 0.50 Punkte je zusätzlichem Punkt IQ_{verbal}
β_{11} $IQ_{verbal\ j}$ (fixed variable)	Zusätzlicher Einfluss der verbalen Intelligenz auf die Änderung der Leistung zum Zeitpunkt i=3 im Vergleich zu Zeitpunkt i = 1. Bsp.: Es sei $\beta_5 = 0.5$. Dann erhöht sich die Vorhersage der Leistungszunahme des Schülers zum Zeitpunkt i = 3 um Vergleich zu Zeitpunkt i = 1 um 0.50 Punkte je zusätzlichem Punkt IQ_{verbal}
$IQ_{figural\text{-}räumlich\ j}$ (Kovariate)	(Grand-mean-)zentrierte figural-räumliche Intelligenz des Schülers j: Bsp.: Es sei 105 die gemessene figural-räumliche Intelligenz des Schülers j und der Durchschnitt der figural-räumlichen Intelligenz sei 99. Dann ist $IQ_{figural\text{-}räumlich\ j} = 6$

β_{12} $IQ_{\text{figural-räumlich j}}$ (fixed variable)	Einfluss der figural-räumlichen Intelligenz auf die Leistung. (Grand-mean-) zentrierte Intelligenz des Schülers j
β_{13} $IQ_{\text{figural-räumlich j}}$ (fixed variable)	Zusätzlicher Einfluss der figural-räumlichen Intelligenz auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1.
β_{14} $IQ_{\text{figural-räumlich j}}$ (fixed variable)	Zusätzlicher Einfluss der figural-räumlichen Intelligenz auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zu Zeitpunkt i = 1.
$IQ_{\text{figural-logisch j}}$ (Kovariate)	(Grand-mean-)zentrierte figural-logische Intelligenz des Schülers j: Bsp.: Es sei 105 die gemessene figural-logische Intelligenz des Schülers j und der Durchschnitt der figural-logischen Intelligenz sei 99. Dann ist $IQ_{\text{figural-logisch j}} = 6$
β_{15} $IQ_{\text{figural-logisch j}}$ (fixed variable)	Einfluss der figural-logischen Intelligenz auf die Leistung. (Grand-mean-)zentrierte figural-logische Intelligenz des Schülers j
β_{16} $IQ_{\text{figural-logisch j}}$ (fixed variable)	Zusätzlicher Einfluss der figural-logischen Intelligenz auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1.
β_{17} $IQ_{\text{figural-logisch j}}$ (fixed variable)	Zusätzlicher Einfluss der figural-logischen Intelligenz auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zu Zeitpunkt i = 1.
Klassengröße_k (Kovariate)	(Grand-mean-)zentrierte Klassengröße der Klasse k.
β_{18} Klassengröße_k (fixed variable)	Einfluss der Klassengröße auf die Leistung.
β_{19} Klassengröße_k (fixed variable)	Zusätzlicher Einfluss der Klassengröße auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1.
β_{20} Klassengröße_k (fixed variable)	Zusätzlicher Einfluss der Klassengröße auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zu Zeitpunkt i = 1.

$Schultyp_k$ (Dummy-Kodierung / Kovariate)	Schultyp der Klasse k. $Schultyp_k = \begin{cases} 1 & falls\ Klasse\ k\ an\ Gymnasium \\ 0 & falls\ Klasse\ k\ an\ IGS \end{cases}$
β_{21} $Schultyp_k$ (fixed variable)	Einfluss des Schultyps auf die Leistung.
β_{22} $Schultyp_k$ (fixed variable)	Zusätzlicher Einfluss des Schultyps auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1.
β_{23} $Schultyp_k$ (fixed variable)	Zusätzlicher Einfluss des Schultyps auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zu Zeitpunkt i = 1.
LES_j (Kovariate)	(Grand-mean-)zentrierte Einschätzung des Lehrerengagement aus Sicht des Schülers (LES) j zum Zeitpunkt i = 1, d.h. vor Beginn der Unterrichtsreihe.
β_{24} LES_j (fixed variable)	Einfluss von LES auf die Leistung des Schülers j.
β_{25} LES_j (fixed variable)	Zusätzlicher Einfluss von LES auf die Änderung der Leistung zum Zeitpunkt i = 2 im Vergleich zu Zeitpunkt i = 1.
β_{26} LES_j (fixed variable)	Zusätzlicher Einfluss von LES auf die Änderung der Leistung zum Zeitpunkt i = 3 im Vergleich zu Zeitpunkt i = 1.

ε_{ij}
(Random Variable)

Fehlerterm. Abweichung der tatsächlich gemessenen Leistung des Schülers j der Klasse k zum Zeitpunkt i von der durch das Modell vorhergesagten Leistung.
ε_{ij} ist für jede Kombination aus i/j Realisation einer (univariat) normalverteilten Zufallsvariablen mit Mittelwert $\mu = 0$ und Standardabweichung $g_i \cdot \sigma_\varepsilon$. Die Standardabweichung ist damit für jeden Zeitpunkt i = 1, 2, 3 unterschiedlich.
Die ε_{ij} sind unabhängig von den Random Interzepten u und p aber untereinander in jedem Individuum korreliert.
D.h.: $(\varepsilon_{1j}, \varepsilon_{2j}, \varepsilon_{3j})$ sind gemeinsam multivariat normalverteilt mit Varianz-Kovarianz-Matrix Σ:

$$\Sigma = \sigma_\varepsilon{}^2 \cdot \begin{pmatrix} 1 & \rho_{12} g_2 & \rho_{13} g_3 \\ \rho_{12} g_2 & g_2^2 & \rho_{23} g_2 g_3 \\ \rho_{13} g_3 & \rho_{23} g_2 g_3 & g_3^2 \end{pmatrix}$$

ϱ_{12} Korrelation zwischen ε_{1j} und ε_{2j}.
ϱ_{13} Korrelation zwischen ε_{1j} und ε_{3j}.
ϱ_{23} Korrelation zwischen ε_{2j} und ε_{3j}.
$\sigma_\varepsilon{}^2$ Varianz von ε_{1j}
$g_2{}^2 \cdot \sigma_\varepsilon{}^2$ Varianz von ε_{2j}
$g_3{}^2 \cdot \sigma_\varepsilon{}^2$ Varianz von ε_{3j}
(man beachte dass $g_1 = 1$ ist / siehe oben)
Damit adressiert der Fehler über die Gewichte g_i die unterschiedlichen Varianzen der Leistungstestergebnisse zu den verschiedenen Messzeitpunkten (Heteroskedastizität). Ferner wird die Korrelation der Messzeitpunkte untereinander berücksichtigt, welche sich aus dem Repeated-Measurement-Design ergeben. Dies ist zwar bereits durch den Individuenparameter p adressiert. Durch diesen ist jedoch die Korrelation der Messzeitpunkte immer gleich, wohingegen die beliebige Korrelation der Fehlerterme höhere Flexibilität bietet.

Man beachte auch die Darstellung:

$$\varepsilon_{ij} = t_1 \varepsilon_{1j} + t_2 \varepsilon_{2j} + t_3 \varepsilon_{3j}$$

Darstellung in der Schreibweise als Hierarchisches Modell:

Messzeitpunkt

$$y_{ijk} = \gamma_{0jk} + \varepsilon_{ij} =$$
$$= \gamma_{0jk} + t_1\,\varepsilon_{1j} + t_2\,\varepsilon_{2j} + t_3\,\varepsilon_{3j}$$

Individuum

$$\gamma_{0jk} = \pi_{0k} + \pi_1 t_2 + \pi_2 t_3 +$$
$$\pi_6\,\text{Notenfaktor}_j + \pi_7 t_2\,\text{Notenfaktor}_j + \pi_8 t_3\,\text{Notenfaktor}_j +$$
$$\pi_9\,IQ_{\text{verbal } j} + \pi_{10} t_2\,IQ_{\text{verbal } j} + \pi_{11} t_3\,IQ_{\text{verbal } j} +$$
$$\pi_{12}\,IQ_{\text{figural-räumlich } j} + \pi_{13} t_2\,IQ_{\text{figural-räumlich } j} + \pi_{14} t_3\,IQ_{\text{figural-räumlich } j} +$$
$$\pi_{15}\,IQ_{\text{figural-logisch } j} + \pi_{16} t_2\,IQ_{\text{figural-logisch } j} + \pi_{17} t_3\,IQ_{\text{figural-logisch } j} +$$
$$\pi_{24}\,LES_j + \pi_{25} t_2\,LES_j + \pi_{26} t_3\,LES_j +$$
$$p_j$$

Klasse

$$\pi_{0k} = \beta_0 + \beta_3\,\text{Bedingung}_k + \beta_4\,\text{Bedingung}_k t_2 + \beta_5\,\text{Bedingung}_k t_3 +$$
$$\beta_{18}\,\text{Klassengröße}_k + \beta_{19}\,\text{Klassengröße}_k t_2 + \beta_{20}\,\text{Klassengröße}_k t_3 +$$
$$\beta_{24}\,\text{Schultyp}_k + \beta_{25}\,\text{Schultyp}_k t_2 + \beta_{26}\,\text{Schultyp}_k t_3 +$$
$$u_k$$

$$\pi_m = \beta_m \text{ für } m = 6\text{-}17,\ 21\text{-}23,\ 27\text{-}29$$

Das Modell kann in Anlehnung an Fahrmeir et al. (2009, S. 259) auch wie folgt dargestellt werden (für Individuen j = 1...525, und Klassen k = 1...21):

$$y_{jk} = \begin{pmatrix} y_{1jk} \\ y_{2jk} \\ y_{3jk} \end{pmatrix} = \boldsymbol{X_{jk}} \cdot \boldsymbol{\beta} + \boldsymbol{U} \cdot \boldsymbol{\gamma}_{jk} + \boldsymbol{\varepsilon}_j$$

Wobei

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_{26})^T \in \mathbb{R}^{27\times 1}$$

(Vektor der festen Effekte)

und

$$\boldsymbol{X_{jk}} = (1 \cdot \boldsymbol{A},\ \text{Bedingung}_\text{j} \cdot \boldsymbol{A},\ \text{Notenfaktor}_\text{j} \cdot \boldsymbol{A},\ \text{IQ}_{\text{verb. j}} \cdot \boldsymbol{A},\ \text{IQ}_{\text{fig.räuml. j}} \cdot \boldsymbol{A},\ \text{IQ}_{\text{fig.log. j}} \cdot \boldsymbol{A},\ \text{Klassengrösse}_\text{k} \cdot \boldsymbol{A},\ \text{Schultyp}_\text{k} \cdot \boldsymbol{A},\ \text{LES}_\text{j} \cdot \boldsymbol{A}) \in \mathbb{R}^{3\times 27}$$

(Designmatrix der festen Effekte)

mit

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

und

$$\mathbf{U} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2} \qquad \text{(Designmatrix für Random Effekte)}$$

$$\boldsymbol{\gamma}_{jk} = \begin{pmatrix} p_j \\ u_k \end{pmatrix} \in \mathbb{R}^{2\times 1} \qquad \text{(Random Interzept je Individuum/Klasse)}$$

$$\boldsymbol{\varepsilon}_j = \begin{pmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \varepsilon_{3j} \end{pmatrix} \in \mathbb{R}^{3\times 1} \qquad \text{(Fehlervektor)}$$

Hierbei gilt für die zufälligen Effekte, dass sie gemeinsam normalverteilt sind:

$$\begin{pmatrix} p_j \\ u_k \\ \varepsilon_{1j} \\ \varepsilon_{2j} \\ \varepsilon_{3j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_p^2 & 0 & & \mathbf{0} & \\ 0 & \sigma_u^2 & & & \\ & & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 \cdot \rho_{12} \cdot g_2 & \sigma_\varepsilon^2 \cdot \rho_{13} \cdot g_3 \\ \mathbf{0} & & \sigma_\varepsilon^2 \cdot \rho_{12} \cdot g_2 & \sigma_\varepsilon^2 \cdot g_2^2 & \sigma_\varepsilon^2 \cdot \rho_{13} \cdot g_2 \cdot g_3 \\ & & \sigma_\varepsilon^2 \cdot \rho_{13} \cdot g_3 & \sigma_\varepsilon^2 \cdot \rho_{13} \cdot g_2 \cdot g_3 & \sigma_\varepsilon^2 \cdot g_3^2 \end{pmatrix} \right)$$

Ferner sind die zufälligen Einflussgrößen unabhängig voneinander:

$$p_1, \ldots, p_{525}, u_1 \ldots, u_{21}, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{525} \text{ unabhängig}$$

Allgemeine Anmerkung zum Verständnis der übrigen Modelle zur Untersuchung der Wirkung des Treatments auf das konzeptuelle Verständnis und die Motivation:

1. Zur Modellbildung wurde jeweils in gleicher Weise vorgegangen: Basierend auf dem Ausgangsmodell wurde jeweils eine Kovariate hinzugenommen und analysiert, ob die jeweilige Kovariate je Messzeitpunkt einen signifikanten Erklärungswert für die abhängige Variable aufweist. Die Ergebnisse dieser Analysen sind in den Tabellen 20 und 21 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com zu finden.
2. In Übereinstimmung mit dem Modell zur Analyse der Treatmentwirkung auf die Physikleistung wurde jeweils ein Mehrebenenmodell gewählt, das sowohl die Klassenstruktur, die Messwiederholung durch die Individuenkomponente als auch die Heteroskedastizität einbezieht und eine unstrukturierte Korrelationsmatrix der Fehlerterme im Individuum zugrunde legt.
3. Des Weiteren wurde für jedes Modell die Intra-Klassenkorrelation (*ICC*) je Untersuchungsebene errechnet und geprüft, ob der *ICC* die Verwendung eines Mehrebenenmodells rechtfertigt oder eine klassische Regressionsanalyse das sparsamere und angemessenere Modell darstellen würde. Zur Erinnerung: gemäß Eid et al. (2011, S. 705) besteht bereits bei zehn Gruppen, einem Siginifkanzniveau von $\alpha = 0.05$ und einer Intra-Klassenkorrelation von $\varrho = 0.10$ das Risiko einer statistischen Fehlentscheidung von 50 %.
4. Die Notation der Modellgleichung der übrigen Modelle kann in analoger Weise dargestellt werden wie die Notation für das Modell zur Analyse der Treatmentwirkung auf die Physikleistung.

2.3.9.3 Erste Hypothese: Wirkung des Treatments auf die Physikleistung

Zur Untersuchung der ersten Hypothese wurde geprüft, ob die Treatmentgruppe im Vergleich zur Kontrollgruppe im Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen im Bereich der Strahlenoptik zum zweiten und dritten Messzeitpunkt (nach der Unterrichtsreihe sowie zwei Monate später) einen höheren Lernzuwachs verzeichnen kann als die Kontrollgruppe. Erfasst wurden Wissen, Problemlösen beim Umgang mit fachspezifischen Repräsentationen durch den Physikleistungstest, der aus repräsentationsbezogenen Aufgaben bestand.

Ein Blick auf die deskriptiven Statistiken (vgl. Tabelle 51) zeigt, dass beide Gruppen annähernd mit dem gleichen Ausgangsniveau starten, in der Postmessung die Treatmentgruppe im Vorteil zu sein scheint, dieser Vorteil jedoch zum Messzeitpunkt des Follow-up Tests nicht erhalten bleibt, so dass beide Gruppen in der letzten Messung ein nahezu identisches Niveau erzielen.

Tabelle 51 Deskriptive Statistiken Leistungstest je Bedingung

	Prätest (*N* = 492)		Posttest (*N* = 484)		Follow-up Test (*N* = 496)	
	TG[a] (*n* = 259)	KG[b] (*n* = 233)	TG[a] (*n* = 252)	KG[b] (*n* = 232)	TG[a] (*n* = 264)	KG[b] (*n* = 232)
M	2.61	2.14	18.21	16.79	13.61	13.36
(SD)	(2.62)	(2.14)	(7.80)	(7.48)	(7.14)	(7.57)
Range	0 - 10.00	0 - 9.00	0 -35.00	0 - 32.50	0 - 28.50	0 - 30.50
Im folgenden Mehrebenenmodell berücksichtigte Daten (N = 443)[c]						
	(*n* = 230)	(*n* = 210)	(*n* = 252)	(*n* = 232)	(*n* = 264)	(*n* = 232)
M	2.63	2.18	18.30	16.70	14.02	13.60
(SD)	(2.66)	(2.17)	(7.80)	(7.46)	(7.00)	(7.54)

Anmerkungen. Maximal erreichbare Punktzahl im Leistungstest: 38 Punkte.
[a]Treatmentgruppe, [b]Kontrollgruppe.
[C]Die Verminderung der Stichprobengröße ergibt sich aus fehlenden Kovariaten.

In den Boxplots (vgl. Abbildung 53) sticht besonders der starke Bodeneffekt zum ersten Messzeitpunkt hervor. Dieser Effekt kann darauf zurückgeführt werden, dass die Schüler zuvor keinerlei Unterricht zu dem Lerninhalt erhalten hatten und entsprechend wenige Aufgaben lösen konnten. Vom Prä- zum Posttest ist ein deutlicher Lernzuwachs in beiden Bedingungen zu verzeichnen. Bis zum dritten Messzeitpunkt haben die Schüler zwar offenbar einige der Lerninhalte vergessen, das

Niveau liegt jedoch deutlich über dem Niveau der Ausgangswerte, so dass auch ein mittelfristiger Lernzuwachs zu erkennen ist.

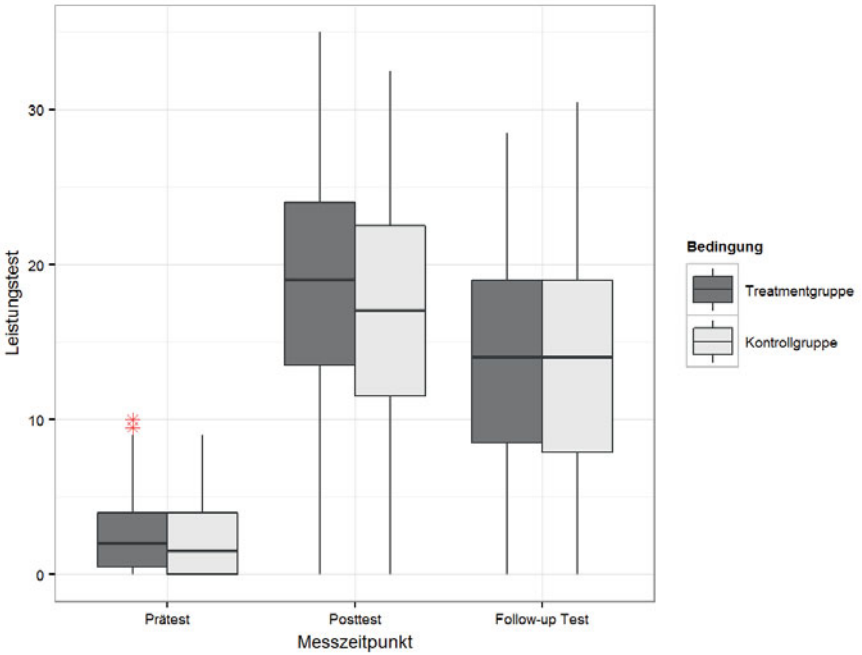


Abbildung 53: Boxplots der Leistungstestergebnisse je Bedingung zu den drei Messzeitpunkten: prä, post und follow-up

In den Verlaufsplots (vgl. Abbildung 54) lässt sich erkennen, dass der Lernzuwachs in beiden Bedingungen nahezu parallel verläuft. Das Leistungsniveau der Gymnasiasten liegt hierbei zu allen drei Messzeitpunkten über dem Niveau der Gesamtschüler.

Basierend auf den zuvor dargestellten Überlegungen wurde nun ein Mehrebenenmodell aufgestellt, das sowohl die Klassenstruktur, die Messwiederholung durch die Individuenkomponente als auch die Heteroskedastizität berücksichtigt und eine unstrukturierte Korrelationsmatrix der Fehlerterme im Individuum zugrunde legt (vgl. Tabelle 52). Entsprechend den vorigen Ergebnissen (vgl. Tabelle 50) wurden folgende Kovariaten berücksichtigt: die erhobenen Schulnoten (Mathematik, Deutsch, Physik), alle drei Subskalen des I-S-T 2000 R und das Lehrerengagement aus Schülersicht vor der Intervention. Des Weiteren wurde der Einfluss des Schultyps und der Klassengröße geschätzt.

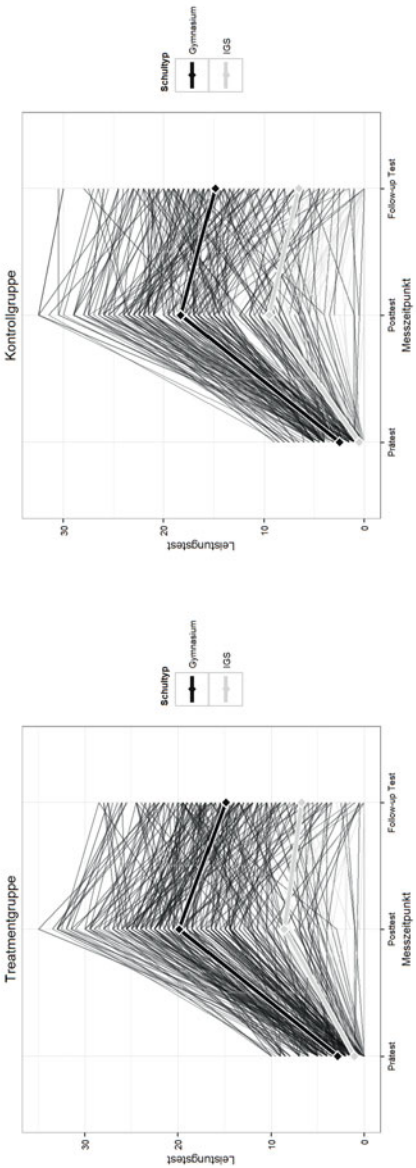


Abbildung 54: Verlaufsplots der Leistungstestergebnisse je Bedingung und Schultyp

Tabelle 52 Ergebnisse des Mehrebenenmodells zu Wissen und Problemlösen bei repräsentationsbezogenen Aufgaben (Leistungstest)

Level (Stichprobengröße)	*Erläuterung*							
Level 1 (*N* = 1263)	Messzeitpunkte							
Level 2 (*N* = 443)	Individuen: Noten, Intelligenz, LES (= Lehrerengagement aus Schülersicht)							
Level 3 (*N*= 21)	Schulklassen: Bedingung Treatment (TG) vs. Kontrollgruppe (KG), Schultyp, Klassengröße							
Fixe Effekte								
Variable	*Erläuterung*	*b*	*(SE)*	*ß*	*(SE)*	$F_{(numDF,\ denDF)}$	*p*
Interzept	Durchschnittliche Leistung: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	1.02	0.61	0.11	0.07	$2.76_{(1,\ 814)}$	.097
Zuwachs (prä – post: t_2-t_1)	Durchschnittlicher Leistungszuwachs: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	9.97	0.87	1.11	0.10	$131.02_{(1,\ 814)}$	< .001
Zuwachs (prä – follow-up: t_3-t_1)	Durchschnittlicher Leistungszuwachs: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	7.29	0.87	0.81	0.10	$70.80_{(1,814)}$	< .001
Bedingung = Treatmentgruppe (TG)	Unterschied zwischen TG und KG zu t_1	0.38	0.46	0.04	0.05	$0.69_{(1,17)}$	.417
Bedingung * Zuwachs prä – post	Zusätzlicher Einfluss Bedingung = TG auf den Zuwachs zwischen t_2 und t_1	0.92	0.59	0.10	0.07	$2.43_{(1,814)}$	.119
Bedingung * Zuwachs prä – follow-up	Zusätzlicher Einfluss der Bedingung = TG auf den Zuwachs zwischen t_3 und t_1	0.19	0.58	0.02	0.06	$0.11_{(1,814)}$	.745
Noten (D, M, PH)	Genereller Einfluss auf Basis von t_1	-0.31	0.08	-0.05	0.01	$15.88_{(1,417)}$	< .001
Noten * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-1.72	0.23	-0.26	0.04	$53.88_{(1,814)}$	< .001
Noten * Prä – Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-1.97	0.24	-0.30	0.04	$70.05_{(1,814)}$	< .001
IQ verbal	Genereller Einfluss auf Basis von t_1	0.04	0.01	0.05	0.01	$15.30_{(1,417)}$	< .001
IQ verbal* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.03	0.03	0.04	0.04	$1.18_{(1,814)}$	.277
IQ verbal Prä – Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.01	0.03	-0.01	0.04	$0.13_{(1,814)}$	.724
IQ figural-räumlich* Ausgangswert	Genereller Einfluss auf Basis von t_1	-0.01	0.01	-0.01	0.01	$0.79_{(1,417)}$	.375

IQ figural-räumlich* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.09	0.03	0.09	0.03	$7.05_{(1,814)}$	.008
IQ figural-räumlich* Prä – Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.03	0.03	0.03	0.03	$0.70_{(1,814)}$	.402
IQ figural-logisch*Ausgangswert	Genereller Einfluss auf Basis von t_1	0.01	0.01	0.01	0.01	$0.57_{(1,417)}$	.450
IQ figural-logisch* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.01	0.03	0.01	0.03	$0.18_{(1,814)}$	.673
IQ figural-logisch* Prä – Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.03	0.03	0.03	0.03	$0.91_{(1,814)}$	.340
Klassengröße	Genereller Einfluss auf Basis von t_1	-0.08	0.08	-0.03	0.03	$0.92_{(1,17)}$	.352
Klassengröße *Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.08	0.11	-0.03	0.04	$0.58_{(1,814)}$	.446
Klassengröße * Prä – Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.07	0.10	0.02	0.04	$0.39_{(1,814)}$	.535
Schultyp * Ausgangswert	Genereller Einfluss Schultyp = Gymnasium auf Basis von t_1	1.39	0.64	0.16	0.07	$4.76_{(1,17)}$	.043
Schultyp * Posttest	Zusätzlicher Einfluss Schultyp = Gymnasium auf den Zuwachs zwischen t_2 und t_1	5.16	0.93	0.57	0.10	$31.05_{(1,814)}$	< .001
Schultyp * Prä – Follow-up Test	Zusätzlicher Einfluss Schultyp = Gymnasium auf den Zuwachs zwischen t_3 und t_1	4.22	0.92	0.47	0.10	$20.96_{(1,814)}$	< .001
LES (vor der Intervention)	Genereller Einfluss auf die Basis von t_1	- 0.02	0.10	0.00	0.01	$0.03_{(1,417)}$	.873
LES (vor der Intervention) * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.29	0.25	0.04	0.03	$1.33_{(1,814)}$	.249
LES (vor der Intervention)* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.27	0.25	-0.04	0.03	$1.22_{(1,814)}$	.271

Zufällige Effekte			
$\sigma_{u\ (Klasse)}$	0.95		
$\sigma_{p\ (Individuum)}$	0.83		
$\sigma_{\varepsilon\ (individuenspezifisch)}$	1.76		
$\rho_{(Messzeitpunkt)}$	(ρ_{12}) 0.12	(ρ_{13}) 0.16	(ρ_{23}) 0.44
$g_{(Messzeitpunkt)}$	(g_1) 1.00	(g_2) 3.36	(g_3) 3.35

Modellvergleich	*Berichtetes Modell*	*Leermodell*[a]
Devianz	7104.25	7358.20
df	35	11

AIC	7380.20			7436.85		
BIC	7174.25			7354.52		
LR-Test (berichtetes Modell versus Leermodell)	253.95, $p < .001$					
Unerklärte Varianz Level 1	(t_1) 4.67	(t_2) 36.26	(t_3) 36.20	(t_1) 5.96	(t_2) 55.51	(t_3) 52.73
Unerklärte Varianz Level 2	25.71			38.07		
Unerklärte Varianz Level 3	1.92			3.53		
Erklärte Varianz des berichteten Modells im Vergleich zum Leermodell						
R^2 Level 1	(t_1) 0.22	(t_2) 0.35	(t_3) 0.31			
R^2 Level 2	0.32					
R^2 Level 3	0.43					
„Wirkung" des Treatments	Anteil aufgeklärter Varianz[b]			Effektstärke[c]		
R^2 Level 1	(t_1) 0.01	(t_2) 0.01	(t_3) < 0.01	$\Delta_{(prä-post)} = 0.15$	$\Delta_{(prä-follow\text{-}up)} = 0.03$	
R^2 Level 2	< 0.01					
R^2 Level 3	0.02					

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen, d.h. das Modell sagt für jeden Schüler voraus, dass seine Leistung zum jeweiligen Messzeitpunkt der mittleren Leistung der Stichprobe entspricht unter Berücksichtigung der individuen-und klassenspezifischen Unterschiede. Man beachte, das Leermodell schätzt folgende 11 Parameter ($df = 11$): Interzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.

[b] Der Anteil aufgeklärter Varianz errechnet sich Snijder und Bosker (1999, S. 99 ff.) folgend, aus:

Level 1 (Messzeitpunkt i): $1-\frac{\sigma_u^2+\sigma_p^2+g_i^2\sigma_\varepsilon^2}{\tilde{\sigma}_u^2+\tilde{\sigma}_p^2+g_i^2\tilde{\sigma}_\varepsilon^2}$

Level 2 (Individuum): $1-\frac{\sigma_u^2+\sigma_p^2+\frac{1}{3}\sum_{i=1}^{3} g_i^2\sigma_\varepsilon^2}{\tilde{\sigma}_u^2+\tilde{\sigma}_p^2+\frac{1}{3}\sum_{i=1}^{3} g_i^2\tilde{\sigma}_\varepsilon^2}$

Level 3 (Klasse): $1-\frac{\sigma_u^2+\frac{1}{h}\left(\sigma_p^2+\frac{1}{3}\sum_{i=1}^{3} g_i^2\sigma_\varepsilon^2\right)}{\tilde{\sigma}_u^2+\frac{1}{h}\left(\tilde{\sigma}_p^2+\frac{1}{3}\sum_{i=1}^{3} g_i^2\tilde{\sigma}_\varepsilon^2\right)}$

h ist hierbei das harmonische Mittel der Klassengrößen $h = 24.60$. Ferner stehen die mit einer Tilde gekennzeichneten Variablen für Variablen aus dem Leermodel und die ungekennzeichneten Variablen für Variablen aus dem berichteten Modell.

[c] Die Effektstärke ergibt sich nach Tymms (2004, S. 56 f.) für dichotome Variablen aus der Differenz der gefragten Mittelwerte (hier mittlerer Lernzuwachs) geteilt durch die gepoolte Standardabweichung (Wurzel der Varianz innerhalb der Gruppen)

$$\Delta_{prä-post} = \frac{\text{Differenz der Klassenmittelwerte des Zuwachses zwischen Treatment- und Kontrollgruppe zu } t_2}{\sqrt{\text{Varianz innerhalb einer Klasse zu } t_2}} = \frac{\beta_4}{\sqrt{\sigma^2_{\text{innerhalb Klasse zu } t_2}}} = \frac{\beta_4}{\sqrt{\sigma_p^2+g_2^2\sigma_\varepsilon^2}}$$

$\Delta_{prä-follow\text{-}up}$ analog

Unter Berücksichtigung der genannten Kovariaten ergaben sich folgende Ergebnisse bezüglich des Lernzuwachses im Post- und im Follow-up Test (Untersuchung der ersten Hypothese): Die Physikleistung bei repräsentationsbezogenen Aufgaben der Schüler war im Prätest zwischen Treatment- und Kontrollgruppe nicht signifikant verschieden ($F_{(1, 17)} = 0.69$, n.s). Betrachtet man den Lernzuwachs vom Präzum Posttest, erzielten Schüler in der Treatmentbedingung einen leicht höheren Lernzuwachs als die Kontrollgruppe. So erreichten die Schüler in der Treatmentgruppe etwa einen Leistungszuwachs, der um einen Leistungspunkt höher liegt ($b = 0.92$) als der Zuwachs der Kontrollgruppe bei einer Maximalpunktzahl von 38 Punkten im Leistungstest, dies entspricht einem Zehntel der Standardabweichung ($\beta = 0.10$). Allerdings handelt es sich um keinen signifikanten Effekt ($F_{(1, 814)} = 2.43$, *n.s.*). Der geringe und auch nicht signifikante Vorteil der Treatmentgruppe bleibt zum Zeitpunkt des Follow-up Tests *nicht* erhalten.

Die Analyse des Anteils aufgeklärter Varianz mit Werten nahe 0 und die berechnete Effektstärke $\Delta < 0.20$ belegen, dass die Bedingung (TG versus KG) keine bzw. kaum eine Relevanz für die Erklärung von Unterschieden im Lernzuwachs zeigt. Von den genannten Kovariaten erwiesen sich die folgenden beiden Kovariaten zu allen drei Messzeitpunkten als hoch signifikant: die Schulnoten, berücksichtigt durch den Notenfaktor und der Schultyp.

Schüler, deren Schulnote um eine Notenstufe besser ist als der Mittelwert der Stichprobe, erzielten über alle Messzeitpunkte hinweg ein Leistungstestergebnis, das knapp einen Drittel Punkt höher lag ($b = -0.31$, $\beta = -0.05$, $p < .001$). Schüler, deren Schulnote um eine Standardabweichung besser ist, erzielten im Schnitt einen Lernzuwachs prä – post, der um ein Viertel einer Standardabweichung höher liegt ($\beta = -0.26$, $p < .001$) und einen Lernzuwachs prä – follow-up, der knapp um ein Drittel einer Standardabweichung ($\beta = -0.30$, $p < .001$) besser ist, als das Ergebnis ihrer Altersgenossen. Dies entspricht einem zusätzlichen Lernzuwachs von knapp 2 Punkten von maximal 38 zu erreichenden Punkten des Leistungstests. Die negativen Vorzeichen ergeben sich aus der Notenskala: höhere Beträge entsprechen schlechteren Bewertungen der Fachleistung.

Gymnasiasten erreichten im Schnitt einen Lernzuwachs im prä – post, der um mehr als eine halbe Standardabweichung ($\beta = 0.57$) höher liegt als der Lernzuwachs der Gesamtschüler und einen Lernzuwachs prä - follow-up, der knapp um eine halbe Standardabweichung ($\beta = 0.47$) von den Ergebnissen der Gesamtschüler abweicht. Dies entspricht einem Vorteil von rund 4 bis 5 Punkten von maximal 38 zu erreichenden Punkten des Leistungstests.

Von den erhobenen Intelligenzkomponenten erweist sich im Gesamtmodell die verbale Intelligenz als über alle Messzeitpunkte hinweg als relevant ($F_{(1, 417)} = 15.30$,

$p < .001$). Sie wirkt sich aber nicht auf den Lernzuwachs aus. Die figural-räumliche Intelligenz beeinflusst signifikant nur den Lernzuwachs prä – post. Der Erklärungsanteil der figural-logischen Intelligenz, die sich in der Einzeltestung als wichtig erwies, wird vermutlich durch die Schulleistung (beispielsweise durch die Noten in Mathematik und Physik) erklärt.

Ein Einfluss der Klassengröße und des Lehrer-Engagements aus Schülersicht ist gemäß des Ergebnisses des Gesamtmodells nicht nachweisbar. Von allen Variablen wirkt sich also der Einfluss des Schultyps am deutlichsten auf den Lernzuwachs prä – post bzw. prä – follow-up aus. Gymnasiasten profitieren mehr von der Intervention als Gesamtschüler. Dies wurde zum Anlass genommen, die Wirkung des Treatments gesondert für Gymnasiasten auszuwerten (siehe Tabelle 53).

Für die gesonderte Auswertung der Stichprobe der Gymnasiasten bestätigt sich, dass auch Schüler an Gymnasien in beiden Bedingungen mit dem gleichen Ausgangsniveau in der Physikleistung starten ($F_{(1, 14)} = 0.71$, *n.s.*). Betrachtet man den Lernzuwachs vom Prä- zum Posttest, erzielten Schüler in der Treatmentbedingung einen höheren Lernzuwachs als die Kontrollgruppe. So erreichten die Schüler in der Treatmentgruppe etwa einen Leistungszuwachs, der knapp um 1.5 Leistungspunkte höher liegt ($b = 1.45$) als die Kontrollgruppe von 38 maximal zu erreichenden Punkten im Leistungstest; dies entspricht $\beta = 0.16$ Standardabweichungen. Im Gegensatz zur Gesamtstichprobe handelt es sich, um einen signifikanten Effekt ($F_{(1,692)} = 5.04$, $p < .05$). Der Anteil aufgeklärter Varianz muss mit Werten nahe 0, ebenso wie die Effektstärke ($\Delta_{\text{prä-post}} = .24$) als gering bewertet werden. Der signifikante Vorteil der Treatmentgruppe bleibt zum Zeitpunkt des Follow-up Tests jedoch auch für die Gymnasiasten alleine *nicht* erhalten. Ebenso wie für die Gesamtstichprobe sind die Schulnoten zu allen drei Messzeitpunkten relevante Einflussfaktoren. Schüler, deren Bewertung der Schulleistung um Notenstufe besser war, erzielten über alle Messzeitpunkte hinweg ein Leistungstestergebnis, das knapp einen halben Punkte höher lag ($b = -0.42$, $\beta = -0.06$, $p < .001$). Zudem kam es im Schnitt zu einem Lernzuwachs prä – post, der knapp um ein Viertel einer Standardabweichung höher lag ($\beta = -0.24$, $p < .001$) und einen Lernzuwachs im Leistungs-Follow-up-Test-Ergebnis, der ebenfalls um in etwas mehr als ein Viertel einer Standardabweichung ($\beta = -0.27$, $p < .001$) höher lag, als der Mittelwert der Stichprobe. Dies entspricht einem zusätzlichen Lernzuwachs von ca. 1.5 Punkten von maximal 38 zu erreichenden Punkten des Leistungstests.

Tabelle 53 Ergebnisse des Mehrebenenmodells zu Wissen und Problemlösen bei repräsentationsbezogenen Aufgaben (Physikleistung), Stichprobe nur Gymnasiasten

Level (Stichprobengröße)	*Erläuterung*						
Level 1 (*N* = 1086)	Messzeitpunkte						
Level 2 (*N* = 376)	Individuen: Noten, Intelligenz, LES (= Lehrerengagement aus Schülersicht)						
Level 3 (*N* = 17)	Schulklassen: Bedingung Treatment (TG) vs. Kontrollgruppe (KG), Schultyp, Klassengröße						
Fixe Effekte							
Variable	*Erläuterung*	*b*	*(SE)*	*ß*	*(SE)*	$F_{(numDF, denDF)}$	*p*
Interzept	Durchschnittliche Leistung: Kontrollgruppe, weiblich, mit durchschnittlichen Werten bezüglich IQ und Noten	2.18	0.42	0.24	0.05	$26.59_{(1,692)}$	< .001
Zuwachs (prä – post: t_2-t_1)	Durchschnittlicher Leistungszuwachs: Kontrollgruppe, weiblich, mit durchschnittlichen Werten bezüglich IQ und Noten	14.66	0.57	1.62	0.06	$658.00_{(1,692)}$	< .001
Zuwachs (prä – follow-up: t_3-t_1)	Durchschnittlicher Leistungszuwachs: Kontrollgruppe, weiblich, mit durchschnittlichen Werten bezüglich IQ und Noten	11.73	0.55	1.30	0.06	$457.66_{(1,692)}$	< .001
Bedingung = Treatmentgruppe (TG)	Unterschied zwischen TG und KG zu t_1	0.47	0.56	0.05	0.06	$0.71_{(1,14)}$	.413
Bedingung * Zuwachs prä – post	Zusätzlicher Einfluss Bedingung = TG auf den Zuwachs zwischen t_2 und t_1	1.45	0.65	0.16	0.07	$5.04_{(1,692)}$	.025
Bedingung * Zuwachs prä – follow-up	Zusätzlicher Einfluss der Bedingung = TG auf den Zuwachs zwischen t_3 und t_1	-0.02	0.62	0.00	0.07	$0.00_{(1,692)}$	.972
Noten (D, M, PH)	Genereller Einfluss auf Basis von t_1	-0.42	0.09	-0.06	0.01	$20.80_{(1,353)}$	< .001
Noten * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-1.62	0.26	-0.24	0.04	$37.65_{(1,692)}$	< .001
Noten * Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-1.84	0.25	-0.27	0.04	$52.84_{(1,692)}$	.001

IQ verbal	Genereller Einfluss auf Basis von t_1	0.05	0.01	0.05	0.01	$16.81_{(1,353)}$	= .001
IQ verbal* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.03	0.03	0.03	0.04	$0.72_{(1,692)}$	.396
IQ verbal* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.02	0.03	-0.03	0.04	$0.51_{(1,692)}$	.475
IQ figural-räumlich* Ausgangswert	Genereller Einfluss auf Basis von t_1	-0.02	0.01	-0.02	0.01	$2.69_{(1,353)}$	.102
IQ figural-räumlich* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.12	0.04	0.12	0.04	$9.81_{(1,692)}$	.002
IQ figural-räumlich* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.05	0.04	0.05	0.04	$2.16_{(1,692)}$	.142
IQ figural-logisch*Ausgangswert	Genereller Einfluss auf Basis von t_1	0.05	0.04	0.01	0.01	$1.12_{(1,692)}$	.291
IQ figural-logisch* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.01	0.01	0.00	0.04	$0.01_{(1,692)}$	.905
IQ figural-logisch* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.00	0.03	0.03	0.04	$0.92_{(1,692)}$	.338
Klassengröße	Genereller Einfluss auf Basis von t_1	0.03	0.03	-0.04	0.03	$1.39_{(1,14)}$	.258
Klassengröße *Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.13	0.11	0.00	0.04	$0.02_{(1,692)}$	.898
Klassengröße *Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.02	0.13	0.14	0.04	$15.70_{(1,692)}$	< .001
LES (vor der Intervention)	Genereller Einfluss auf Basis von t_1	-0.87	0.62	0.00	0.02	$0.02_{(1,353)}$	.877
LES (vor der Intervention) * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.02	0.11	0.02	0.04	$0.31_{(1,692)}$	.577
LES (vor der Intervention)* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.15	0.27	-0.05	0.04	$2.30_{(1,692)}$	.130

Zufällige Effekte

$\sigma_{u\ (Klasse)}$	1.05
$\sigma_{p\ (Individuum)}$	0.93
$\sigma_{\varepsilon\ (individuenspezifisch)}$	1.78

ρ *(Messzeitpunkt)*	(ρ_{12}) 0.14		(ρ_{13}) 0.19	(ρ_{23}) 0.44		
g *(Messzeitpunkt)*	(g_1) 1.00		(g_2) 3.36	(g_3) 3.24		
Modellvergleich	*Berichtetes Modell*			*Leermodell*[a]		
Devianz	6072.56			6261.87		
df	35			11		
AIC	6142.556			6283.872		
BIC	6317.215			6338.765		
LR-Test (berichtetes Modell versus Leermodell)	189.32, $p < .001$					
Unerklärte Varianz Level 1	(t_1) 5.15	(t_2) 37.39	(t_3) 35.31	(t_1) 6.25	(t_2) 48.80	(t_3) 48.59
Unerklärte Varianz Level 2	25.95			34.54		
Unerklärte Varianz Level 3	2.11			2.97		
Erklärte Varianz des berichteten Modells im Vergleich zum Leermodell						
R^2 Level 1	(t_1) 0.18	(t_2) 0.23	(t_3) 0.27			
R^2 Level 2	0.25					
R^2 Level 3	0.30					
„Wirkung" des Treatments	Anteil aufgeklärter Varianz[b]			Effektstärke		
R^2 Level 1	(t_1) 0.01	(t_2) 0.02	(t_3) < 0.01	$\Delta_{(prä-post)} = 0.24$	$\Delta_{(prä-follow\text{-}up)} < 0.01$	
R^2 Level 2	0.01					
R^2 Level 3	0.03					

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen, d.h. das Modell sagt für jeden Schüler voraus, dass seine Leistung zum jeweiligen Messzeitpunkt der mittleren Leistung der Stichprobe entspricht unter Berücksichtigung der individuen-und klassenspezifischen Unterschiede. Man beachte, das Leermodell schätzt folgende 11 Parameter ($df = 11$): Inzerzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.
[b] Der Anteil aufgeklärter Varianz errechnet sich wie in Tabelle 52 angegeben nach Snijder und Bosker (1999, S. 99 ff.)
Für das harmonische Mittel der Klassengrößen ergibt sich $h_{Gymnasium} = 25.20$.

Von den erhobenen Intelligenzkomponenten erweist sich für die Stichprobe der Gymnasiasten ebenfalls die verbale Intelligenz als relevant ($F_{(1, 353)} = 16.81$, $p < .001$). Die figural-räumliche Intelligenz beeinflusst wie zu erwarten ebenfalls den Lernzuwachs prä – post ($F_{(1, 692)} = 9.81$, $p < .01$). Ebenfalls relevant war die Klassengröße, wobei Schüler in größeren Klassen einen höheren Lernzuwachs prä – follow-up erzielten ($\beta = 0.14$, $F_{(1, 692)} = 15.70$, $p < .001$).

2.3.9.4 Zweite Hypothese: Wirkung des Treatments auf das konzeptuelle Verständnis

Zur Untersuchung der zweiten Hypothese wurde geprüft, ob die Treatmentgruppe im Vergleich zur Kontrollgruppe bezüglich des konzeptuellen Verständnisses zum zweiten und dritten Messzeitpunkt (nach der Intervention sowie zwei Monate später) einen höheren Lernzuwachs verzeichnen kann als die Kontrollgruppe. Das konzeptuelle Verständnis wurde durch den Konzepttest Strahlenoptik erhoben. Ein Blick auf die deskriptiven Statistiken zeigt, dass beide Gruppen mit dem annähernd gleichen Ausgangsniveau starten, nach der Unterrichtsreihe ihr konzeptuelles Verständnis verbessern und diese Verbesserung auch erhalten bleibt (vgl. Tabelle 54).

Tabelle 54 Deskriptive Statistiken Konzepttest je Bedingung

	Prätest (N = 491)		Posttest (N = 480)		Follow-up Test (N = 486)	
	TG[a]	KG[b]	TG[a]	KG[b]	TG[a]	KG[b]
	(n = 260)	(n = 231)	(n = 248)	(n = 232)	(n = 258)	(n = 228)
M	7.50	7.41	10.73	11.20	10.91	11.18
(SD)	(3.24)	(3.23)	(4.70)	(5.29)	(4.64)	(4.86)
Range	0 - 18.00	0 - 16.00	0 - 21.00	0 - 22.00	0 - 22.00	0 - 22.00
Im folgenden Mehrebenenmodell berücksichtigte Daten (N = 443)[c]						
	(n = 231)	(n = 210)	(n = 211)	(n = 200)	(n = 217)	(n = 194)
M	7.49	7.49	10.79	11.28	10.91	11.39
(SD)	(3.20)	(3.10)	(4.74)	(5.27)	(4.61)	(4.83)

Anmerkung. Maximal erreichbare Punktzahl im Konzepttest: 22 Punkte.
[a]Treatmentgruppe, [b]Kontrollgruppe.
[c]Die verminderte Stichprobengröße ergibt sich aus fehlenden Kovariaten.

An den Boxplots (vgl. Abbildung 55) zeigt sich, dass Schüler in der Kontrollbedingung zu allen drei Messzeitpunkten leicht im Vorteil zu sein scheinen.

Während zum ersten Messzeitpunkt von Schülern in beiden Bedingungen im Schnitt etwa ein Drittel der Maximalpunktzahl erreicht wird, steigt dieser Wert zum zweiten und dritten Messzeitpunkt auf knapp die Hälfte der maximal erreichbaren Punktzahl. Dieses Niveau bleibt zum dritten Messzeitpunkt in beiden Bedingungen in etwa erhalten.

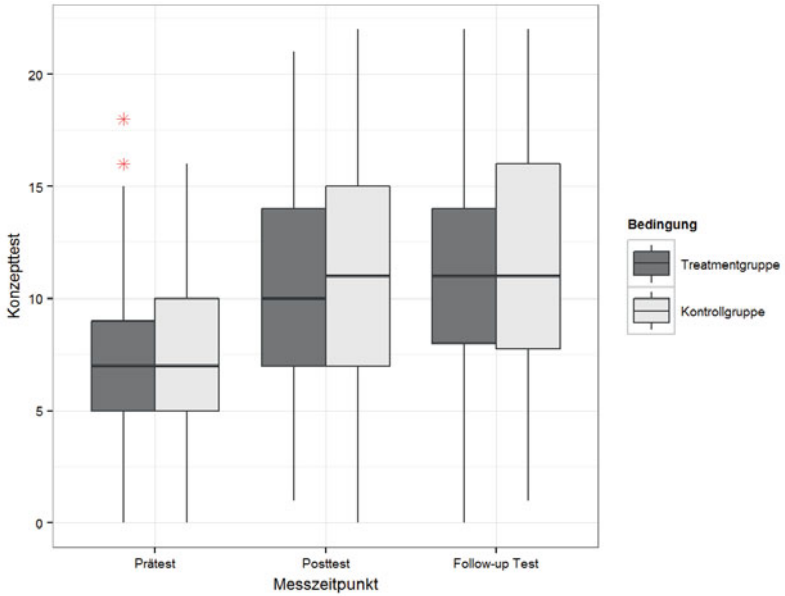


Abbildung 55: Boxplots der Konzepttestergebnisse je Bedingung zu den drei Messzeitpunkten: prä, post und follow-up

Die Verlaufsplots legen nahe, dass sich das konzeptuelle Verständnis in beiden Bedingungen ähnlich entwickelt (vgl. Abbildung 56). Auch im Konzepttest schneiden die Gymnasiasten im Mittel zu allen drei Messzeitpunkten besser ab als die Gesamtschüler.

Ebenso wie im Modell zum Leistungstest, erweist sich auch für den Konzepttest jede der infrage kommenden Variablen außer der Geschlechtszugehörigkeit zu mindestens einem der drei Zeitpunkte als relevant. Entsprechend wurden alle infrage kommenden Variablen außer der Geschlechtszugehörigkeit im Gesamtmodell berücksichtigt. Die ICCs des Modells zur Untersuchung des konzeptuellen

Verständnisses bestätigten, dass die Verwendung eines hierarchischen linearen Modells ($ICC_{(Level\text{-}Schulklasse)} = 0.25$, $ICC_{(Level\text{-}Individuen)} = 0.51$) die angemessene Analyseform darstellte. Die Berücksichtigung der Heteroskedastizizät verbesserte im Hinblick auf die Informationskriterien *AIC* und *BIC* das Modell, daher wurde auch hier mathematisch das gleiche Ausgangsmodell gewählt wie im Leistungstest. Die Ergebnisse des Modells zur Analyse der Wirkung der Intervention auf das konzeptuelle Verständnis finden sich in Tabelle 55.

Unter Berücksichtigung der genannten Kovariaten ergaben sich folgende Ergebnisse bezüglich des Lernzuwachses im Post- und im Follow-up Test (Untersuchung der zweiten Hypothese): Schüler in beiden Bedingungen starten mit dem gleichen Ausgangsniveau bezüglich des konzeptuellen Verständnisses ($F_{(1, 17)} = 0.03$, *n.s.*). Betrachtet man den Lernzuwachs, erzielten Schüler in der Treatmentbedingung weder prä – post ($F_{(1, 802)} = 1.85$), noch prä – follow-up ($F_{(1, 802)} = 1.47$) einen signifikant höheren Lernzuwachs als die Kontrollgruppe (*n.s.*). Werte nahe 0 für den Anteil aufgeklärter Varianz (der sich auf das Treatment zurückführen lässt) und die geringe Effektstärke Δ unterstützen den Befund, dass keine Unterschiede feststellbar sind.

Von den genannten Kovariaten erwiesen sich lediglich die Schulnoten (berücksichtigt durch den Notenfaktor durch das Ergebnis der PCA, siehe Kapitel 2.3.6.2 Ergebnisse zu vorherigen Schulleistungen) im Gesamtmodell zu allen drei Messzeitpunkten als hoch signifikant. Schüler, deren Bewertung der Schulleistung um eine Notenstufe besser war, erzielten über alle Messzeitpunkte hinweg ein Testergebnis, das um einen Drittel Punkt besser war (b = -0.33). Ausgedrückt in Standardabweichungen bedeutet dies: Weichen die Noten um eine Standardabweichung ab, ergibt sich ein Unterschied im Leistungstest von β = -0.10. Schüler, deren Bewertung der Schulleistung um eine Standardabweichung besser war, erzielten darüber hinaus im Schnitt noch zusätzlich einen Lernzuwachs prä – post, der ebenfalls um knapp ein Drittel einer Standardabweichung höher lag (β = -0.27) sowie einen Lernzuwachs vom Prä- zum Follow-up Test, der etwa um ein Viertel einer Standardabweichung (β = -0.26) höher lag als der Mittelwert. Dies entspricht bezüglich des Lernzuwachses in beiden Fällen einem Vorteil von knapp einem Punkt von maximal 22 zu erreichenden Punkten des Konzepttests.

Von den erhobenen Intelligenzkomponenten erweisen sich im Gesamtmodell die verbale Intelligenz und das figural-räumliche Schlussfolgern prinzipiell als relevant ($p < .01$). Weicht der gemessene Wert im I-S-T 2000 R um eine Standardabweichung ab, so erzielen die Schüler ein Ergebnis im Konzepttest, das um (etwa) eine Zehntel Standardabweichungen höher liegt als der Mittelwert: $\beta_{IQ\text{-}figural\text{-}räumlich * Ausgangswert} = 0.08$, $\beta_{\text{IQ-verbal * Ausgangswert}} = 0.10$).

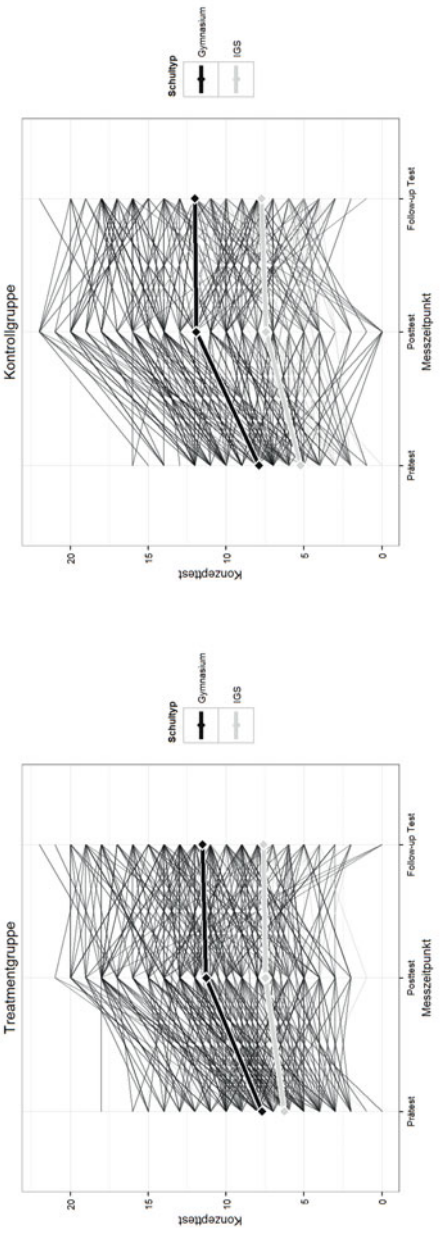


Abbildung 56: Verlaufsplots der Konzepttestergebnisse je Bedingung und Schultyp

Tabelle 55 Ergebnisse des Mehrebenenmodells zum konzeptuellen Verständnis in der Strahlenoptik

Level (Stichprobengröße)	*Erläuterung*						
Level 1 (*N* = 1263)	Messzeitpunkte						
Level 2 (*N* = 443)	Individuen: Noten, Intelligenz, LES (= Lehrerengagement aus Schülersicht)						
Level 3 (*N* = 21)	Schulklassen: Bedingung Treatment- (TG) vs. Kontrollgruppe (KG), Schultyp, Klassengröße						
Fixe Effekte							
Variable	*Erläuterung*	*b*	*(SE)*	*β*	*(SE)*	$F_{(numDF, denDF)}$	*p*
Interzept	Durchschnittliches Verständnis: Kontrollgruppe, weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	6.76	0.73	1.45	0.16	$86.71_{(1,802)}$	< .001
Zuwachs (prä – post: t_2-t_1)	Durchschnittlicher Verständniszuwachs: Kontrollgruppe, weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	3.08	0.62	0.66	0.13	$25.04_{(1,802)}$	< .001
Zuwachs (prä – follow-up: t_3-t_1)	Durchschnittlicher Verständniszuwachs: Kontrollgruppe, weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	2.72	0.59	0.58	0.13	$21.10_{(1,802)}$	< .001
Bedingung = Treatmentgruppe (TG)	Unterschied zwischen TG und KG zu t_1	0.09	0.54	0.02	0.12	$0.03_{(1,17)}$	.873
Bedingung * Zuwachs prä – post	Zusätzlicher Einfluss der Bedingung = TG auf den Zuwachs zwischen t_2 und t_1	-0.57	0.42	-0.12	0.09	$1.85_{(1,802)}$	.174
Bedingung * Zuwachs prä – follow-up	Zusätzlicher Einfluss der Bedingung = TG auf den Zuwachs zwischen t_3 und t_1	-0.49	0.40	-0.10	0.09	$1.47_{(1,802)}$	.226
Noten (D, M, PH)	Genereller Einfluss auf Basis von t_1	-0.33	0.11	-0.10	0.03	$9.46_{(1,417)}$	.002
Noten * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.91	0.16	-0.27	0.05	$31.28_{(1,802)}$	<.0001
Noten * Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.88	0.16	-0.26	0.05	$29.01_{(1,802)}$	<.0001
IQ verbal	Genereller Einfluss auf Basis von t_1	0.04	0.01	0.10	0.03	$9.45_{(1,417)}$	.002
IQ verbal* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.01	0.02	0.03	0.05	$0.27_{(1,802)}$	.603
IQ verbal* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.01	0.02	0.02	0.05	$0.17_{(1,802)}$	.684
IQ figural-räumlich* Ausgangswert	Genereller Einfluss auf Basis von t_1	0.04	0.02	0.08	0.03	$6.59_{(1,417)}$	.011
IQ figural-räumlich* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.03	0.02	-0.06	0.05	$1.72_{(1, 802)}$	.190
IQ figural-räumlich* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.06	0.02	-0.11	0.05	$5.68_{(1,802)}$	.017

IQ figural-logisch*Ausgangswert	Genereller Einfluss auf Basis von t_1			0.02	0.01	0.05	0.03	$2.59_{(1,417)}$	.108
IQ figural-logisch* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1			-0.01	0.02	-0.02	0.05	$0.22_{(1,802)}$	.637
IQ figural-logisch* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1			-0.01	0.02	-0.03	0.05	$0.37_{(1,802)}$	.543
Klassengröße	Genereller Einfluss auf Basis von t_1			0.07	0.09	0.04	0.06	$0.55_{(1,17)}$	.467
Klassengröße *Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1			0.11	0.07	0.07	0.05	$2.34_{(1,802)}$	.127
Klassengröße *Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1			0.10	0.07	0.07	0.05	$2.19_{(1,802)}$	.140
Schultyp * Ausgangswert	Genereller Einfluss des Schultyps = Gymnasium auf Basis von t_1			0.68	0.76	0.15	0.16	$0.80_{(1,17)}$	.384
Schultyp * Posttest	Zusätzlicher Einfluss des Schultyps = Gymnasium auf den Zuwachs zwischen t_2 und t_1			0.68	0.65	0.15	0.14	$1.10_{(1,802)}$	.296
Schultyp * Follow-up Test	Zusätzlicher Einfluss des Schultyps = Gymnasium auf den Zuwachs zwischen t_3 und t_1			1.26	0.63	0.27	0.14	$4.02_{(1,802)}$	.045
LES (vor der Intervention)	Genereller Einfluss auf Basis von t_1			0.18	0.13	0.05	0.03	$2.03_{(1,417)}$	.155
LES (vor der Intervention) * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1			0.31	0.18	0.08	0.05	$3.08_{(1,802)}$	.080
LES (vor der Intervention)* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1			-0.06	0.17	-0.02	0.04	$0.11_{(1,802)}$	.735
Zufällige Effekte									
$\sigma_{u\ (Klasse)}$	1.08								
$\sigma_{p\ (Individuum)}$	1.55								
$\sigma_{\varepsilon\ (individuenspezifisch)}$	2.15								
$\rho_{(Messzeitpunkt)}$	(ρ_{12}) -0.01			(ρ_{13}) 0.04		(ρ_{23}) 0.25			
$g_{(Messzeitpunkt)}$	(g_1) 1.00			(g_2) 1.66		(g_3) 1.61			
Modellvergleich	*Berichtetes Modell*			*Leermodell*[a]					
Devianz	6597.94			6764.41					
df	35			11					
AIC	6667.94			6786.411					
BIC	6847.88			6842.964					
LR-Test (berichtetes Modell versus Leermodell)	166.47***								
Unerklärte Varianz Level 1	(t_1) 8.20	(t_2) 16.31	(t_3) 15.59	(t_1) 10.83		(t_2) 21.49		(t_3) 21.01	

Unerklärte Varianz Level 2	13.37			17.78	
Unerklärte Varianz Level 3	1.66			3.76	
Erklärte Varianz des berichteten Modells im Vergleich zum Leermodell					
R^2 Level 1	(t_1) 0.24	(t_2) 0.24	(t_3) 0.26		
R^2 Level 2	0.25				
R^2 Level 3	0.57				
„Wirkung“ des Treatments	Anteil aufgeklärter Varianz[b]			Effektstärke	
R^2 Level 1	(t_1) 0.01	(t_2) 0.03	(t_3) 0.02	$\Delta_{(prä-post)}$ = -0.15	$\Delta_{(prä-follow\text{-}up)}$ = -0.13
R^2 Level 2	< 0.01				
R^2 Level 3	< 0.01				

* $p < .05$, **$p < .01$, ***$p < .001$

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen. Das Modell sagt also für jeden Schüler voraus, dass seine Leistung zum jeweiligen Messzeitpunkt dem mittleren konzeptuellen Verständnis der Stichprobe entspricht - unter Berücksichtigung der individuen-und klassenspezifischen Unterschiede. Man beachte, das Leermodell schätzt folgende 11 Parameter (df = 11): Interzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.

[b] Der Anteil aufgeklärter Varianz errechnet sich wie in Tabelle 52 angegeben nach Snijder und Bosker (1999, S. 99 ff.).

Für das harmonische Mittel der Klassengrößen ergibt sich h = 24.60

Der Schultyp zeigt im Gesamtmodell lediglich auf, dass Gymnasiasten mittelfristig (prä – follow-up) das Gelernte besser behalten als Gesamtschüler ($F_{(1, 802)} = 4.02, p < .05$). Sie erzielten durchschnittlich einen Lernzuwachs, der um knapp ein Drittel einer Standardabweichung ($\beta = 0.27$) prä – follow-up höher liegt als der Zuwachs der Gesamtschüler. Dies entspricht einem Vorteil von etwas mehr als einem Punkt von maximal 22 zu erreichenden Punkten des Konzepttests. Zu den anderen Messzeitpunkten konnten keine signifikanten Unterschiede im Erwerb des konzeptuellen Verständnisses zwischen Gymnasiasten und Gesamtschülern festgestellt werden. Ein Einfluss der Klassengröße und des Lehrer-Engagements aus Schülersicht (LES) ist im Gesamtmodell nicht mehr nachweisbar.

2.3.9.5 Dritte Hypothese: Wirkung der Treatmentvariation auf das konzeptuelle Verständnis

Zur Untersuchung der dritten Hypothese wurde geprüft, ob Schüler, welche ein variiertes Treatment erhalten hatten, zum zweiten und dritten Messzeitpunkt (nach der Intervention sowie zwei Monate später) in Bezug auf das konzeptuelle Verständnis einen höheren Lernzuwachs verzeichnen können, als Schüler, welche das reguläre Treatment erhalten hatten bzw. als Schüler in der Kontrollbedingung.

Die deskriptiven Statistiken zeigen, dass alle Gruppen in etwa mit dem gleichen Ausgangsniveau starten, wobei die Schüler in der variierten Treamentbedingung die höchsten Werte im Prätest erzielten.

Nach der Unterrichtsreihe verbessern Schüler in allen Bedingungen ihr konzeptuelles Verständnis und diese Verbesserung bleibt auch in allen Bedingungen erhalten (vgl. Tabelle 56).

Bei dem Blick auf die Boxplots deutet sich an, dass Schüler in der variierten Treatmentbedingung im Prä- und Posttest leicht im Vorteil zu sein scheinen (vgl. Abbildung 57). Im Posttest erreichen Schüler in der variierten Treatmentbedingung im Schnitt die höchsten Punktzahlen und auch den höchsten Zuwachs im Vergleich zum Prätest. Im Follow-up Test scheinen Schüler in der Kontrollbedingung am besten abzuschneiden und im Vergleich zu den vorherigen Messzeitpunkten den höchsten Lernzuwachs zu erzielen. Zu prüfen ist, ob diese Unterschiede signifikant sind.

Tabelle 56 Deskriptive Statistiken Konzepttest je Bedingung

	Prätest (*N* = 491)			Posttest (*N* = 480)			Follow-up Test (*N* = 486)		
	TG var.[a]	TG reg.[b]	KG[c]	TG var.[a]	TG reg.[b]	KG[b]	TG var.[a]	TG reg.[b]	KG[b]
	(*n* = 91)	(*n* = 169)	(*n* = 231)	(*n* = 91)	(*n* = 157)	(*n* = 232)	(*n* = 92)	(*n* = 166)	(*n* = 228)
M	7.83	7.28	7.41	12.01	10.04	11.20	10.68	11.05	11.18
(SD)	3.32	3.11	(3.23)	4.96	4.45	(5.29)	5.09	4.33	(4.86)
Range	0 - 15.00	1 - 18	0 – 16.00	1 - 20.00	1 - 21.00	0 - 22.00	0 - 20.00	2 - 22.00	0 - 22.00
Im folgenden Mehrebenenmodell berücksichtigte Daten (N = 443)[c]									
	(*n* = 87)	(*n* = 144)	(*n* = 210)	(*n* = 80)	(*n* = 131)	(*n* = 200)	(*n* = 80)	(*n* = 137)	(*n* = 194)
M	7.83	7.28	7.49	12.01	10.04	11.28	10.68	11.05	11.39
(SD)	3.32	3.11	(3.10)	4.96	4.45	(5.27)	5.09	4.33	(4.83)

Anmerkung: Maximal erreichbare Punktzahl im Konzepttest: 22 Punkte,
[a]variiertes Treatment, [b]reguläres Treatment, [c]Kontrollgruppe.
[C]Die Verminderung der Stichprobengröße ergibt sich aus fehlenden Kovariaten.

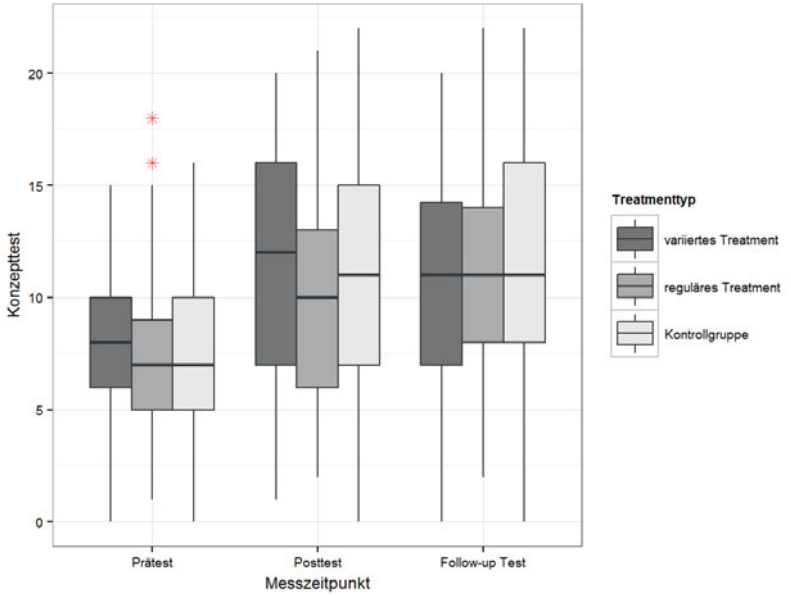


Abbildung 57: Boxplots der Konzepttestergebnisse je Bedingung zu den drei Messzeitpunkten: prä, post und follow-up

Um die Ergebnisse direkt mit den Ergebnissen des vorigen Mehrebenenmodells zum konzeptuellen Verständnis vergleichen zu können (vgl. Tabelle 55), wurden die gleichen Kovariaten wie in dem vorherigen Modell berücksichtigt. Die Ergebnisse sind in Tabelle 57 wieder gegeben.

Die Ergebnisse des Modells (vgl. Tabelle 57) zeigen, dass Schüler in den verschiedenen Bedingungen mit dem gleichen Ausgangsniveau bezüglich des konzeptuellen Verständnisses starten: Dies gilt sowohl für den Vergleich zwischen regulärem Treatment und Kontrollgruppe ($F_{(1, 16)} = 0.00$, *n.s.*), für den Vergleich zwischen variiertem Treatment und Kontrollgruppe ($F_{(1, 16)} = 0.12$, *n.s.*) als auch für den Vergleich zwischen regulärem Treatment und variiertem Treatment ($F_{(1, 16)} = 0.10$, *n.s.*).

Betrachtet man den Lernzuwachs erzielten Schüler in der Kontrollgruppe prä – post ($F_{(1, 800)} = 4.72$, $p = .030$) einen signifikant höheren Lernzuwachs als in der regulären Treatmentbedingung, dies gilt jedoch nicht für den Lernzuwachs prä – follow-up ($F_{(1, 800)} = .013$, n.s.). Das variierte Treamtent erwies sich als signifikant besser als das reguläre Treatment prä – post ($F_{(1, 800)} = 4.05$, $p = .045$), dieser Unterschied blieb prä – follow-up ($F_{(1, 800)} = 2.27$, n.s.) jedoch nicht erhalten. Zwischen variierem Treatment- und Kontrollgruppe ergaben sich weder ein Unterschied prä – post ($F_{(1, 800)} = 0.10$, n.s.) noch prä – follow-up ($F_{(1, 800)} = 3.68$, n.s.).

Die in den Boxplots ersichtlichen deskriptiven Vorteile zugunsten des variierten Treatments (vgl. Abbildung 57) werden also bestätigt. Werte nahe 0 für den Anteil aufgeklärter Varianz und die geringe Effektstärke Δ belegen jeodch, dass die genannten Unterschiede praktisch wenig bedeutsam sind (vgl. Tabelle 57).

Von den genannten Kovariaten erwiesen sich lediglich die Schulnoten (berücksichtigt durch den Notenfaktor als Ergebnis der PCA, siehe Kapitel 2.3.6.2 Ergebnisse zu vorherigen Schulleistungen) im Gesamtmodell zu allen drei Messzeitpunkten als hoch signifikant. Je besser die Schulnoten waren, desto höher fiel die Testleistung aus. Schüler, deren Bewertung der Schulleistung um eine Notenstufe besser war, erzielten über alle Messzeitpunkte hinweg im Schnitt ein Testergebnis, das um einen drittel Punkt höher lag ($b = 0.33$, $\beta = -0.10$). Schüler, deren Bewertung der Schulleistung um Standardabweichung besser war, erzielten darüber hinaus im Schnitt noch zusätzlich einen Lernzuwachs prä – post, der ebenfalls um knapp ein Drittel einer Standardabweichung höher lag ($\beta = -0.27$) sowie einen Lernzuwachs vom Prä- zum Follow-up Test, der etwa um ein Viertel einer Standardabweichung ($\beta = -0.26$) höher ausfiel, als der Mittelwert der Gesamtstichprobe. Dies entspricht bezüglich des Lernzuwachses in beiden Fällen einem Vorteil von knapp einem Punkt von maximal 22 zu erreichenden Punkten des Konzepttests.

Tabelle 57 Ergebnisse des Mehrebenenmodells zum Vergleich der Wirkung des regulären Treatments mit der Treatmentvariation auf das konzeptuelle Verständnis in der Strahlenoptik

Level (Stichprobengröße)	**Erläuterung**						
Level 1 (*N* = 1263)	Messzeitpunkte						
Level 2 (*N* = 443)	Individuen: Noten, Intelligenz, LES (= Lehrerengagement aus Schülersicht)						
Level 3 (*N*= 21)	Schulklassen: Bedingung Treatment- (TG) vs. Kontrollgruppe (KG), Schultyp, Klassengröße						
Fixe Effekte							
Variable	*Erläuterung*	*b*	*(SE)*	*β*	*(SE)*	$F_{(numDF,\ denDF)}$	*p*
Interzept	Durchschnittliches Verständnis: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	6.74	0.73	1.44	0.16	$86.00_{(1,800)}$	< .001
Zuwachs (prä – post: t_2-t_1)	Durchschnittlicher Verständniszuwachs: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	2.94	0.62	0.63	0.13	$22.72_{(1,800)}$	< .001
Zuwachs (prä – follow-up: t_3-t_1)	Durchschnittlicher Verständniszuwachs: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	2.84	0.60	0.61	0.13	$22.68_{(1,800)}$	< .001
Bedingung = Treatmentgruppe regulär (TG)	Unterschied zwischen regulärem Treatment und Kontrollgruppe zum Zeitpunkt 1	0.01	0.61	-0.01	0.13	$0.00_{(1,16)}$	.997
Bedingung * Zuwachs prä – post	Unterschied im Lernzuwachs zwischen t_2 und t_1 zwischen regulärem Treatment und Kontrollgruppe	-1.03	0.48	-0.22	0.10	$4.72_{(1,800)}$	.030
Bedingung * Zuwachs prä – follow-up	Unterschied im Lernzuwachs zwischen t_3 und t_1 zwischen regulärem Treatment und Kontrollgruppe	-0.16	0.46	-0.04	0.10	$0.13_{(1,800)}$	.719
Bedingung = Treatmentgruppe variiert (TG)	Unterschied zwischen regulärem Treatment und Treatmentvariation zum Zeitpunkt 1	0.25	0.78	0.05	0.17	$0.10_{(1,16)}$	.751
Bedingung * Zuwachs prä – post	Unterschied im Lernzuwachs zwischen t_2 und t_1 zwischen regulärem Treatment und Treatmentvariation	1.21	060	0.26	0.13	$4.02_{(1,800)}$	.045
Bedingung * Zuwachs prä – follow-up	Unterschied im Lernzuwachs zwischen t_3 und t_1 zwischen regulärem Treatment und Treatmentvariation	-0.86	0.58	0.12	0.12	$2.27_{(1,800)}$	.132
Bedingung = Treatment regulär + Treatment variiert	Unterschied zwischen Treatmentvariation und Kontrollgruppe zum Zeitpunkt 1 (Summe der obigen Effekte)	0.25	0.73	0.05	0.16	$0.12_{(1,16)}$	.737

Bedingung * Zuwachs prä – post	Unterschied im Lernzuwachs zwischen t_2 und t_1 zwischen Treatmentvariation und Kontrollgruppe zum (Summe der obigen Effekte)	0.18	0.56	0.04	0.12	$0.10_{(1,800)}$	.747
Bedingung * Zuwachs prä – follow-up	Unterschied im Lernzuwachs zwischen t_3 und t_1 zwischen Treatmentvariation und Kontrollgruppe (Summe der obigen Effekte)	-1.03	0.54	-0.22	0.12	$3.68_{(1,800)}$	.056
Noten (D, M, PH)	Genereller Einfluss auf Basis von t_1	-0.33	0.11	-0.10	0.03	$9.44_{(1,417)}$	.002
Noten * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.92	0.16	-0.27	0.05	$31.82_{(1,800)}$	.722
Noten * Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.88	0.16	-0.26	0.05	$29.03_{(1,800)}$	.601
IQ verbal	Genereller Einfluss auf Basis von t_1	0.04	0.01	0.10	0.03	$9.36_{(1,417)}$	.011
IQ verbal* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.01	0.02	0.02	0.05	$0.13_{(1,800)}$	.126
IQ verbal* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.01	0.02	0.02	0.05	$0.27_{(1,800)}$	.026
IQ figural-räumlich* Ausgangswert	Genereller Einfluss auf Basis von t_1	0.04	0.02	0.08	0.03	$6.52_{(1,417)}$	.109
IQ figural-räumlich* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.04	0.02	-0.07	0.05	$2.35_{(1,800)}$	.580
IQ figural-räumlich* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.05	0.02	-0.10	0.04	$4.99_{(1,800)}$	.592
IQ figural-logisch*Ausgangswert	Genereller Einfluss auf Basis von t_1	0.02	0.01	0.05	0.05	$2.58_{(1,417)}$	.472
IQ figural-logisch* Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.01	0.02	-0.03	0.04	$0.31_{(1,800)}$	.167
IQ figural-logisch* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.01	0.02	-0.02	0.06	$0.29_{(1,800)}$	.121
Klassengröße	Genereller Einfluss auf Basis von t_1	0.07	0.07	-0.04	0.05	$10.54_{(1,16)}$	.369
Klassengröße *Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.10	0.76	0.07	0.05	$1.92_{(1,800)}$	.193
Klassengröße *Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	0.11	0.66	0.07	0.05	$2.41_{(1,800)}$	.078
Schultyp * Ausgangswert	Genereller Einfluss des Schultyps = Gymnasium auf Basis von t_1	0.70	0.63	0.15	0.16	$0.86_{(1,16)}$	.157

Schultyp * Posttest	Zusätzlicher Einfluss des Schultyps = Gymnasium auf den Zuwachs zwischen t_2 und t_1	0.86	0.03	0.18	0.14	$1.70_{(1,800)}$	.142
Schultyp * Follow-up Test	Zusätzlicher Einfluss des Schultyps = Gymnasium auf den Zuwachs zwischen t_3 und t_1	1.12	0.04	0.24	0.14	$3.12_{(1,800)}$	.881
LES (vor der Intervention)	Genereller Einfluss auf Basis von t_1	0.18	0.13	0.05	0.03	$2.01_{(1,417)}$	< .001
LES (vor der Intervention) * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	0.26	0.18	0.07	0.05	$2.16_{(1,800)}$	< .001
LES (vor der Intervention)* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.03	0.17	-0.01	0.04	$0.02_{(1,800)}$	< .001

Zufällige Effekte						
$\sigma_{u\ (Klasse)}$	1.07					
$\sigma_{p\ (Individuum)}$	1.57					
$\sigma_{\varepsilon\ (individuenspezifisch)}$	2.14					
$\rho_{(Messzeitpunkt)}$	(ρ_{12}) -0.01			(ρ_{13}) 0.03		(ρ_{23}) 0.26
$g_{(Messzeitpunkt)}$	(g_1) 1.00			(g_2) 1.65		(g_3) 1.61
Modellvergleich	*Berichtetes Modell*			*Leermodell*[a]		
Devianz	6597.94			6261.87		
df	38			11		
AIC	6662.72			6786.41		
BIC	6858.08			6842.96		
LR-Test (berichtetes Modell versus Leermodell)	177.69***					
Unerklärte Varianz Level 1	(t_1) 8.20	(t_2) 16.12	(t_3) 15.52	(t_1) 10.83	(t_2) 21.49	(t_3) 21.01
Unerklärte Varianz Level 2	13.28			17.78		
Unerklärte Varianz Level 3	1.65			3.76		
Erklärte Varianz des berichteten Modells im Vergleich zum Leermodell						
R^2 Level 1	(t_1) 0.24	(t_2) 0.25	(t_3) 0.26			
R^2 Level 2	0.25					
R^2 Level 3	0.56					

„Wirkung“ des Treatments	Anteil aufgeklärter Varianz[b]			Effektstärke		
R^2 Level 1	(t_1) < 0.01	(t_2) 0.01	(t_3) 0.01	reguläres Treatment vs. Kontrollgruppe	$\Delta_{(prä-post)}$ = - 0.27	$\Delta_{(prä-follow\text{-}up)}$ = - 0.04
R^2 Level 2	0.01			variiertes Treatment vs. Kontrollgruppe	$\Delta_{(prä-post)}$ = 0.05	$\Delta_{(prä-follow\text{-}up)}$ = - 0.27
R^2 Level 3	0.01					

* $p < .05$, **$p < .01$, ***$p < .001$

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen. Das Modell sagt für jeden Schüler voraus, dass seine Leistung zum jeweiligen Messzeitpunkt dem mittleren konzeptuellen Verständnis der Stichprobe entspricht - unter Berücksichtigung der individuen-und klassenspezifischen Unterschiede. Man beachte, das Leermodell schätzt folgende 11 Parameter (df = 11): Interzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.

[b] Der Anteil aufgeklärter Varianz errechnet sich wie in Tabelle 52 angegeben nach Snijder und Bosker (1999, S. 99 ff.).

Für das harmonische Mittel der Klassengrößen ergibt sich h = 24.60.

Wie zu erwarten war, unterschieden sich die Ergebnisse bezüglich des Einflusses der Kovariaten kaum im Vergleich zu dem Mehrebenenmodell zum konzeptuellen Verständnis, welches nicht zwischen regulärem und variiertem Treatment differenziert.

Von den erhobenen Intelligenzkomponenten erwies sich ebenfalls das verbale Schlussfolgern als relevant: Wich der gemessene Wert im I-S-T 2000 R um eine Standardabweichung ab, so erzielen die Schüler ein Ergebnis im Konzepttest, das um knapp ein Zehntel einer Standardabweichung höher lag ($\beta = 0.10, p < .05$).

Der Schultyp erwies sich in diesem Modell nunmehr nicht mehr als signifikant. Die Klassengröße und das Lehrer-Engagement aus Schülersicht spielten ebenfalls keine signifikante Rolle.

2.3.9.6 Vierte Hypothese: Wirkung des Treatments auf die Schülermotivation

Zur Untersuchung der Wirkung des Unterrichts in der Treatment- und in der Kontrollbedingung auf die Schülermotivation wurde untersucht, ob sich die Motivation im Verlauf in den beiden Bedingungen unterscheidet. Die deskriptiven Statistiken ergeben, dass beide Gruppen mit dem gleichen Ausgangsniveau starten, im Verlauf des Schuljahrs die Motivation jedoch in beiden Bedingungen sinkt und sich zum zweiten und dritten Messzeitpunkt auf einem etwas niedrigeren Niveau stabilisiert (vgl. Tabelle 58).

Tabelle 58 Deskriptive Statistiken Motivation je Bedingung

	Prätest (*N* = 489)		Posttest (*N* = 477)		Follow-up Test (*N* = 495)	
	TG[a]	KG[b]	TG[a]	KG[b]	TG[a]	KG[b]
	(*n* = 259)	(*n* = 230)	(*n* = 253)	(*n* = 224)	(*n* = 264)	(*n* = 231)
M	3.59	3.60	3.40	3.43	3.43	3.42
(SD)	(0.80)	(0.78)	(0.88)	(0.85)	(0.86)	(0.90)
Range	1.08 - 5.63	1.33 – 5.36	1.00 – 6.00	1.21 – 5.58	1.00 – 6.00	1.00 – 5.71
Im folgenden Mehrebenenmodell berücksichtigte Daten (N = 443)[c]						
	(*n* = 231)	(*n* = 212)	(*n* = 219)	(*n* = 193)	(*n* = 222)	(*n* = 222)
	3.57	3.60	3.43	3.43	3.40	3.38
	(0.82)	(0.81)	(0.91)	(0.88)	(0.87)	(0.91)

Anmerkung. Skala von 1 (niedrige Ausprägung) - 6 (hohe Ausprägung)

[a]Treatmentgruppe, [b]Kontrollgruppe

[c]Die verminderte Stichprobenzahl ergibt sich aus fehlenden Kovariaten.

Augenscheinlich sind in den Boxplots (vgl. Abbildung 58) kaum Unterschiede zwischen den Bedingungen zu erkennen. Wie sich bereits an den Mittelwerten erkennen ließ, sinkt das Motivationsniveau nach der Unterrichtsreihe leicht ab und stabilisiert sich. D.h. das etwas verminderte Motivationsniveau nach der Intervention bleibt zum dritten Messzeitpunkt in beiden Bedingungen in etwa erhalten.

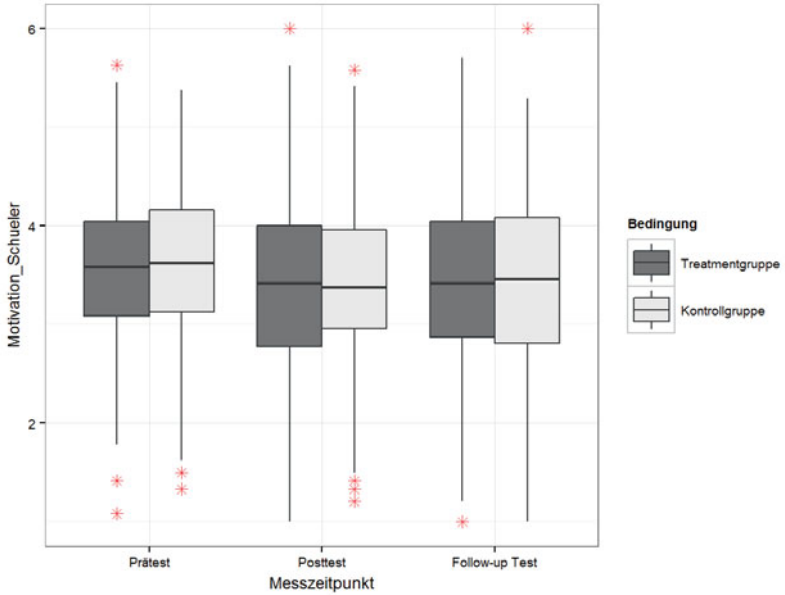


Abbildung 58: Boxplots zur Schülermotivation je Bedingung

Ein interessantes Detailergebnis zeigt sich im Hinblick auf den Motivationsverlauf je Schultyp in der Treatmentbedingung (vgl. Abbildung 59). Im Gegensatz zum Motivationsverlauf der Gymnasiasten, deutet sich für die Gesamtschüler tendentiell ein u-förmiger Verlauf an: D.h. die Motivation der Schüler erreicht ihren Tiefpunkt unmittelbar nach der Unterrichtsreihe und steigt zwei Monate später nach dem regulären Physikunterricht etwas an, wobei zum letzten Messzeitpunkt in etwa das gleiche Motivationsniveau wie an den Gymnasien erreicht wird.

Dabei gilt es zu prüfen, ob der Unterschied im Motivationsverlauf prä – post zwischen den Schultypen Gymnasium und Gesamtschule signifikant ist und ob in Bezug auf den Schultyp ATI-Effekt vorliegt.

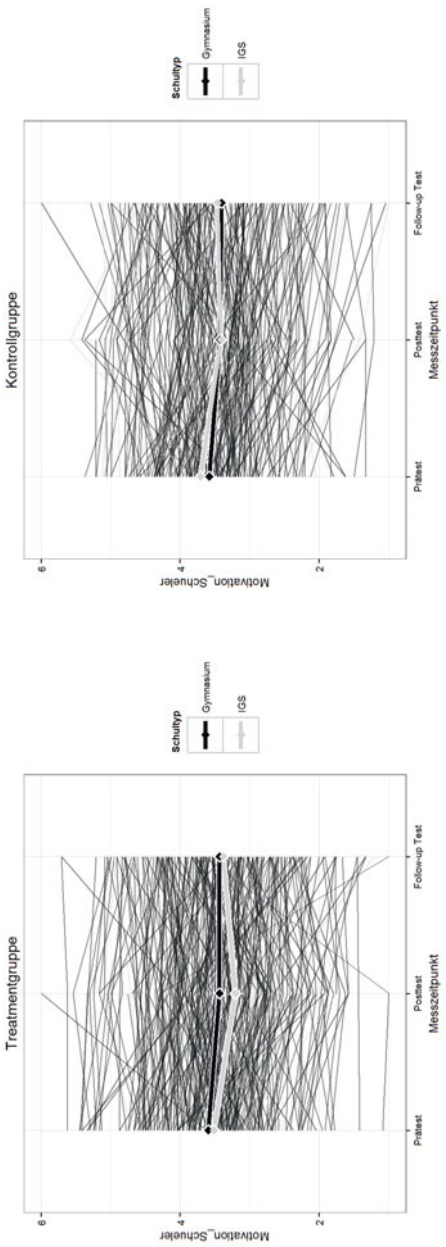


Abbildung 59: Verlaufsplots zur Schülermotivation je Bedingung und Schultyp

Auf Basis der Ergebnisse der Einzeltestung relvanter Kovariaten wurden die Variablen: Schulleistungen (Noten), das figural-logische Schlussfolgern, das Lehrerengagement aus Sicht der Schüler (LES) und die Geschlechtszugehörigkeit der Schultyp in dem Mehrebenenmodell zur Analyse des Motivationsverlaufs berücksichtigt.

Der in den Verlaufsplots deskriptiv ersichtliche Unterschied im Motivationsverlauf prä – post zwischen Gymnasiasten und Gesamtschülern ist in der Einzeluntersuchung der Kovariaten knapp nicht signifikant ($F_{(1, 931)} = 3.01$, $p = .083$). Um dennoch den deskriptiv ersichtlichen Unterschieden nachzugehen, wurde die Variable „Schultyp" in das Gesamtmodell aufgenommen.

Die Werte des ICC's des Modells zur Untersuchung Schülermotivation sprechen für die Verwendung eines hierarchischen linearen Modells anstelle einer klassichen Regressionsanalyse ($ICC_{(Klasse)} = 0.14$, $ICC_{(Individuen)} = 0.73$). Gemessen an den Informationskritrien *AIC* und *BIC* verbesserte die Berücksichtigung der Heteroskedastizität das Modell.

Unter Berücksichtigung der genannten Kovariaten ergaben sich folgende Ergebnisse zur Entwicklung der Schülermotivation (vgl. Tabelle 59): Schüler in beiden Bedingungen starten mit dem gleichen Ausgangsniveau ($F_{(1, 18)} = 0.72$, n.s.). Betrachtet man die Veränderung im weiteren Verlauf des Schuljahrs nach der Unterrichtsreihe, ist in der Entwicklung der Schülermotivation weder prä – post ($F_{(1, 816)} = 0.06$, n.s.), noch prä – follow-up ($F_{(1, 816)} = 2.19$, n.s.) eine signifikant unterschiedliche Entwicklung zwischen Treatment- und Kontrollgruppe zu beobachten. Die Werte für den Anteil der aufgeklärten Varianz durch das Treatment und die Effektstärke unterstreichen diesen Befund.

Von den berücksichtigten Kovariaten erwiesen sich folgende Variablen im Gesamtmodell als relevant: die Fachnoten, die Geschlechtszugehörigkeit, der Schultyp und das Lehrerengagement aus Schülersicht (LES).

Schüler, deren Bewertung der Schulleistung um eine Notenstufe höher liegt, stuften ihre Motivation im Schnitt über alle Messzeitpunkte hinweg um 4 Punkte höher ein ($b = -0.19$, $\beta = -0.30$, $p < .001$), allerdings sank die Schülermotivation „besserer" Schüler auch signifikant stärker von der Prä- zur Postmessung. Dabei handelt es sich jedoch nur um einen kleinen Effekt: Weichen die Fachnoten um eine Standardabweichung nach oben ab, sinkt die Schülermotivation prä – post um $\beta = 0.03$ Standardabweichungen. Dies entspricht einem Rückgang von etwa 0.09 Punkten auf einer Skala von 1-6 Punkten.

Tabelle 59 Ergebnisse des Mehrebenenmodells zur Schülermotivation

Level (Stichprobengröße)	*Erläuterung*						
Level 1 (*N = 1273*)	Messzeitpunkte						
Level 2 (*N = 443*)	Individuen: Noten, Intelligenz, LES (= Lehrerengagement aus Schülersicht)						
Level 3 (*N = 21*)	Schulklassen: Bedingung Treatment- (TG) vs. Kontrollgruppe (KG), Schultyp, Klassengröße						
Fixe Effekte							
Variable	*Erläuterung*	*b*	*(SE)*	*ß*	*(SE)*	$F_{(numDF,\ denDF)}$	*p*
Interzept	Durchschnittlicher Wert: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	3.72	0.13	4.29	0.14	$875.21_{(1,816)}$	< .001
Zuwachs (prä – post: t_2-t_1)	Durchschnittliche Veränderung: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	-0.28	0.09	-0.32	0.10	$10.67_{(1,816)}$	=.001
Zuwachs (prä – follow-up: t_3-t_1)	Durchschnittlicher Veränderung: Kontrollgruppe, Geschlechtszugehörigkeit = weiblich, IGS, mit durchschnittlichen Werten bezüglich IQ und Noten	-0.29	0.09	-0.33	0.11	$9.64_{(1,816)}$	.002
Bedingung = Treatmentgruppe (TG)	Unterschied zwischen TG und KG zu t_1	-0.08	0.09	-0.09	0.11	$0.72_{(1,18)}$	.407
Bedingung * Zuwachs prä – post	Zusätzlicher Einfluss der Bedingung = TG auf die Veränderung zwischen t_2 und t_1	0.01	0.06	0.02	0.07	$0.06_{(1,816)}$	.815
Bedingung * Zuwachs prä – follow-up	Zusätzlicher Einfluss der Bedingung = TG auf die Veränderung zwischen t_3 und t_1	0.09	0.06	0.11	0.07	$2.19_{(1,816)}$	.139
Noten (D, M, PH)	Genereller Einfluss auf Basis der Schätzung zu t_1	-0.19	0.02	-0.30	0.04	$60.62_{(1,418)}$	< .001
Noten * Posttest	Zusätzlicher Einfluss auf die Veränderung zwischen t_2 und t_1	0.06	0.02	0.09	0.03	$7.72_{(1,816)}$	.006
Noten * Follow-up Test	Zusätzlicher Einfluss auf die Veränderung zwischen t_3 und t_1	0.01	0.03	0.01	0.04	$0.09_{(1,816)}$	.766
IQ figural-logisch*Ausgangswert	Genereller Einfluss auf Basis der Schätzung zu t_1	-0.00	0.03	-0.02	0.04	$0.21_{(1,418)}$	.646
IQ figural-logisch* Posttest	Zusätzlicher Einfluss auf die Veränderung zwischen t_2 und t_1	0.00	0.03	0.04	0.03	$1.26_{(1,816)}$	.262
IQ figural-logisch* Follow-up Test	Zusätzlicher Einfluss auf die Veränderung zwischen t_3 und t_1	0.01	0.03	0.07	0.04	$3.43_{(1,816)}$	.064
Geschlecht * Ausgangswert	Einfluss Geschlechtszugehörigkeit = männlich zu t_1	0.32	0.06	0.37	0.07	$26.04_{(1,418)}$	< .001

Geschlecht * Posttest	Zusätzlicher Einfluss der Geschlechtszugehörigkeit = männlich auf die Veränderung zwischen t_2 und t_1	0.02	0.06	0.02	0.07	$0.09_{(1,816)}$	.764
Geschlecht * Follow-up Test	Zusätzlicher Einfluss der Geschlechtszugehörigkeit = männlich auf die Veränderung zwischen t_3 und t_1	-0.08	0.06	-0.09	0.07	$1.41_{(1,816)}$	.236
Schultyp * Ausgangswert	Genereller Einfluss des Schultyps = Gymnasium auf Basis der Schätzung zu t_1	-0.30	0.13	-0.35	0.15	$5.74_{(1,18)}$*	.028
Schultyp * Posttest	Zusätzlicher Einfluss des Schultyps = Gymnasium auf die Veränderung zwischen t_2 und t_1	0.13	0.08	0.15	0.10	$2.43_{(1,816)}$	.120
Schultyp * Follow-up Test	Zusätzlicher Einfluss des Schultyps = Gymnasium auf die Veränderung zwischen t_3 und t_1	0.09	0.09	0.10	0.11	$0.96_{(1,816)}$	.327
LES (vor der Intervention)	Genereller Einfluss auf Basis der Schätzung zu t_1	0.29	0.03	0.41	0.04	$101.20_{(1,418)}$	<.001
LES (vor der Intervention) * Posttest	Zusätzlicher Einfluss auf die Veränderung zwischen t_2 und t_1	-0.01	0.02	-0.02	0.03	$0.31_{(1,816)}$	.576
LES (vor der Intervention)* Follow-up Test	Zusätzlicher Einfluss auf die Veränderung zwischen t_3 und t_1	-0.09	0.03	-0.13	0.04	$12.26_{(1,816)}$	< .001

Zufällige Effekte						
$\sigma_{u\ (Klasse)}$	0.16					
$\sigma_{\rho\ (Individuum)}$	0.53					
$\sigma_{\varepsilon\ (individuenspezifisch)}$	0.32					
ρ *(Messzeitpunkt)*	(ρ_{12}) 0.13		(ρ_{13}) 0.10		(ρ_{23}) 0.43	
g *(Messzeitpunkt)*	(g_1) 1.00		(g_2) 1.63		(g_3) 1.84	
Modellvergleich	*Berichtetes Modell*			*Leermodell*[a]		
Devianz	2237.29			2414.37		
df	26			11		
AIC	2295.29			2436.36		
BIC	2444.61			2493.00		
LR-Test (berichtetes Modell versus Leermodell)	170.17***					
Unerklärte Varianz Level 1	(t_1) 0.41	(t_2) 0.58	(t_3) 0.66	(t_1) 0.65	(t_2) 0.76	(t_3) 0.83
Unerklärte Varianz Level 2	0.55			0.75		
Unerklärte Varianz Level 3	0.05			0.14		

Erklärte Varianz des berichteten Modells im Vergleich zum Leermodell

R^2 Level 1	(t_1) 0.37	(t_2) 0.23	(t_3) 0.20		
R^2 Level 2	0.26				
R^2 Level 3	0.67				
„Wirkung" des Treatments	Anteil aufgeklärter Varianz[b]			Effektstärke	
R^2 Level 1	(t_1) < 0.01	(t_2) < 0.01	(t_3) < 0.01	$\Delta_{(prä-post)} = 0.01$	$\Delta_{(prä-follow\text{-}up)} = 0.11$
R^2 Level 2	< 0.01				
R^2 Level 3	0.02				

* $p < .05$, ** $p < .01$, *** $p < .001$

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen. Das Modell sagt für jeden Schüler voraus, dass die Motivation zum jeweiligen Messzeitpunkt den Mittelwert der Motivation in der Stichprobe entspricht - unter Berücksichtigung der individuen-und klassenspezifischen Unterschiede. Man beachte, das Leermodell schätzt folgende 11 Parameter (df = 11): Inzerzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.

[b] Der Anteil aufgeklärter Varianz errechnet sich wie in Tabelle 52 angegeben nach Snijder und Bosker (1999, S. 99 ff.).

Für das harmonische Mittel der Klassengrößen ergibt sich $h = 24.60$.

Jungen stuften ihre Motivation über alle Messzeitpunkte hinweg um 0.32 Punkte signifikant höher ein als Mädchen ($F_{(1,418)} = 26.04$, $p < .001$). Dies entspricht etwas mehr als einem Drittel einer Standardabweichung ($\beta = 0.37$). Bezüglich der weiteren Motivationsentwicklung konnten jedoch keine signifikanten Unterschiede zwischen Mädchen und Jungen festgestellt werden.

Gymnasiasten schätzten ihre Motivation über alle Messzeitpunkte hinweg um etwa $b = -0.30$ Punkte niedriger ein als Gesamtschüler, was in etwa einem Drittel einer Standardabweichung entspricht ($\beta = -0.35$, $F_{(1, 18)} = 5.74$, $p < .05$). Bezüglich der weiteren Motivationsentwicklung konnten jedoch keine signifikanten Unterschiede zwischen Gymnasiasten und Gesamtschülern festgestellt werden.

Schultyp und Geschlechtszugehörigkeit wirken sich also in etwa gleichem Maß auf die Schülermotivation über alle Messzeitpunkte hinweg aus, tragen jedoch nicht dazu bei, Unterschiede in Entwicklungsverläufen der Motivation aufzuklären.

Das Lehrer-Engagement aus Schülersicht wirkt sich über alle Messzeitpunkte hinweg höchst signifikant auf die Schülermotivation aus ($F_{(1, 418)} = 101.20$, $p < .001$). Weicht das subjektiv individuell bewertete Engagement des Lehrers um eine Standardabweichung nach oben ab, liegt die selbsteingeschätzte Motivation um $\beta = 0.41$ Standardabweichungen höher. Dies entspricht $b = 0.29$ Punkten auf einer Skala von 1 bis 6 Punkten. Allerdings sank die Motivation von Schülern, die das Lehrerengagement höher einstuften, auch signifikant stärker von der Prä- zur Follow-up-Messung (nicht jedoch von der Prä- zur Postmessung). Dabei handelt es sich jedoch nur um einen kleineren Effekt: weicht die Bewertung des Lehrerengagements um eine Standardabweichung nach oben ab, sinkt die Schülermotivation prä – follow-up um $\beta = -0.13$ Standardabweichungen. Dies entspricht einem Rückgang von $b = -0.09$ Punkten auf einer Skala von 1 bis 6 Punkten.

Die figural-logische Intelligenz spielt gemäß der Ergebnisse des Gesamtmodells nunmehr keine signifikante Rolle für die Motivationsentwicklung weder generell über alle Messzeitpunkte hinweg noch im weiteren Schuljahresverlauf.

2.3.9.7 Fünfte Hypothese: Vergleich inhaltlich aufeinander bezogener Studien

Abschließend wurden die Daten des Konzepttests mit den Daten der Studie von Scheid (2013) verglichen. Ziel war es zu analysieren, ob sich der in der Pilotstudie gefundene starke Effekt für die Überlegenheit der Thematisierung von Schülervorstellungen mit einer größeren Stichprobe, wenn auch unter variierten Randbedingungen, bestätigen lässt. Zudem wurde überprüft, ob die Thematisierung von Schülervorstellungen im Unterricht nicht nur unmittelbar nach der Intervention,

sondern auch mittelfristig (ca. zwei Monate später) das konzeptuelle Verständnis fördern kann.

Gemeinsamkeiten beider Studien: In beiden Stichproben wurde exakt das gleiche Testinstrument verwendet. Für die im Folgenden berichteten deskriptiven Statistiken und für das Mehrebenenmodell zur Prüfung signifikanter Unterschiede im Lernzuwachs wurde der Gesamtscore aus den gleichen Items ermittelt.

Je eine Treatment- und eine Kontrollgruppe wurden in beiden Studien je Schule vom gleichen Lehrer unterrichtet. Schüler beider Studien besuchten die 7. oder 8. Klassenstufe einer weiterführenden Schule im Bundesland Rheinland-Pfalz (M = 13 [Jahre], SD = 8 [Monate]). Lerninhalt in beiden Studien war die Bildentstehung bei der Sammellinse gemäß Lehrplan des Bundeslandes Rheinland-Pfalz im Fach Physik der Klassenstufen 7 bzw. 8.

In beiden Studien fanden drei Datenerhebungen statt: vor der Unterrichtsreihe, nach der Unterrichtsreihe, die in beiden Studien sechs bis sieben Schulstunden umfasste, sowie einige Wochen nach der Unterrichtsreihe (follow-up). Die zeitlichen Abstände unterscheiden sich lediglich zum dritten Messzeitpunkt, während die Schüler der Stichprobe „Ko“ sechs Wochen später getestet wurden, fand die Follow-up Erhebung der Stichprobe „SV“ acht bis neun Wochen später statt.

Unterschiede beider Studien: Während in dieser Studie („SV“) sowohl in der Treatment- als auch in der Kontrollgruppe Schülervorstellungen adressiert worden sind, lag der Schwerpunkt in der Studie („Ko“) von Scheid (2013) auf der Förderung der Kohärenz im Umgang mit Repräsentationen. Entsprechend spielte die Thematisierung von Schülervorstellungen keine explizite Rolle im Unterricht der Treatment- und der Kontrollbedingung. Eine ausführliche Darstellung des Unterrichts in der Treatment- und Kontrollbedingung kann in der Dissertationsschrift von Scheid (2013) nachgelesen werden. Die unterschiedlichen Werte des Medians des konzeptuellen Verständnisses vor der Intervention (siehe Boxplots, Abbildung 60) resultieren daraus, dass an der Studie „SV“ (dunkelgrau) auch Gesamtschüler teilnahmen, an der Vergleichsstichprobe „Ko“ (hellgrau) jedoch nur Gymnasiasten. Weitere Angaben zur Stichprobe der Studie „Ko“ finden sich in der Dissertationsschrift von Scheid (2013).

Vergleichbarkeit der Daten von Treatment- und Kontrollgruppe innerhalb der beiden Studien: Da innerhalb der Stichproben in keiner der beiden Studien signifikante Unterschiede zwischen Treatment- und Kontrollbedingung festzustellen waren, wurde keine weitere Differenzierung nach Bedingung innerhalb der Stichproben der Studien vorgenommen. So bestanden weder Unterschiede im Ausgangsniveau vor der Unterrichtsreihe noch im späteren Lernzuwachs zwischen Treatment- und Kontrollgruppe in der Stichprobe „SV“ (siehe Kapitel 2.3.9.4).

Zweite Hypothese: Wirkung des Treatments auf das konzeptuelle Verständnis) und auch keine Unterschiede im Ausgangsniveau vor der Unterrichtsreihe und im Lernzuwachses zwischen Treatment- und Kontrollgruppe in der Stichprobe „Ko“ (siehe Scheid, 2013, S. 173, 174). In den folgenden Analysen wurde also Unterricht unter Berücksichtigung von Schülervorstellungen („SV“) mit Unterricht verglichen, in dem das nicht der Fall war („Ko“).

Im Hinblick auf die deskriptiven Statistiken zeigt sich (vgl. Tabelle 60), dass Schüler in beiden Stichproben mit annähernd gleichen Kenntnissen starten, wobei die Stichprobe „SV“im Schnitt einen Mittelwert, der um einen halben bzw. um einen ganzen Punkt auf einer Skala von 0 - 22 nach unten abweicht. Diese leicht unterschiedlichen Ausgangswerte zu Beginn der Studien resultieren vermutlich daraus, dass an der Studie „SV“ auch Gesamtschüler teilnahmen, an der Vergleichsstichprobe „Ko“ (Scheid, 2013) jedoch nur Gymnasiasten. Zu prüfen ist, ob dieser Anfangsunterschied signifikant ist, zumal bezüglich der Variablen Schultyp keine signifikanten Unterschiede für die Stichprobe „SV“ festgestellt werden konnten (siehe Kapitel 2.3.9.4 Zweite Hypothese: Wirkung des Treatments auf das konzeptuelle Verständnis).

Tabelle 60 Deskriptive Statistiken – Konzepttest Stichprobenvergleich (*N* = 988)

	Prätest (*N* = 938)		Posttest (*N* = 889)		Follow-up Test (*N* = 926)	
	„SV“[a] (*n* = 480)	„Ko“[b] (*n* = 444)	„SV“[a] (*n* = 472)	„Ko“[b] (*n* = 384)	„SV“[a] (*n* = 476)	„Ko“[b] (*n* = 436)
M	7.44	8.08	10.96	9.41	11.04	9.57
(SD)	3.23	3.01	5.01	3.68	4.75	3.54
Range	0 - 18.00	0 - 18.00	0 - 22.00	0 - 20.00	0 - 22.00	0 - 22.00

Anmerkung. Maximal erreichbare Punktzahl im Konzepttest: 22 Punkte.
[a]Schülervorstellungen, [b]Kohärenz
[c] Im Stichprobenvergleich liegen von allen teilnehmenden Schüler vollständige Kovariaten vor, daher entspricht die berichtete Stichprobengröße der Datenbasis des folgenden Mehrebenenmodells. Die Schülerleistungen zählten zur Datenbasis, wenn das jeweilige Individuum zu mindestens einem der Messzeitpunkte den Test bearbeitete und alle Fachnoten bekannt waren.

Kovariaten vor, daher entspricht die berichtete Stichprobengröße der Datenbasis des folgenden Mehrebenenmodells. Die Schülerleistungen zählten zur Datenbasis, wenn das jeweilige Individuum zu mindestens einem der Messzeitpunkte den Test bearbeitete und alle Fachnoten bekannt waren.

An den Boxplots wird ersichtlich (vgl. Abbildung 60), dass Schüler beider Stichproben ihr konzeptuelles Verständnis prä – post und prä – follow-up ver-

bessern, wobei der Zuwachs der Stichprobe „SV“ den Zuwachs der Stichprobe „Ko“ übertrifft. Zum Zeitpunkt der Follow-up Messung bleibt das Wissensniveau in beiden Stichproben in etwa erhalten. Zudem zeigt sich besonders deutlich, dass die Berücksichtigung von Schülervorstellungen (Stichprobe „SV“) zu einem noch deutlicheren Lernzuwachs im konzeptuellen Verständnis führt.

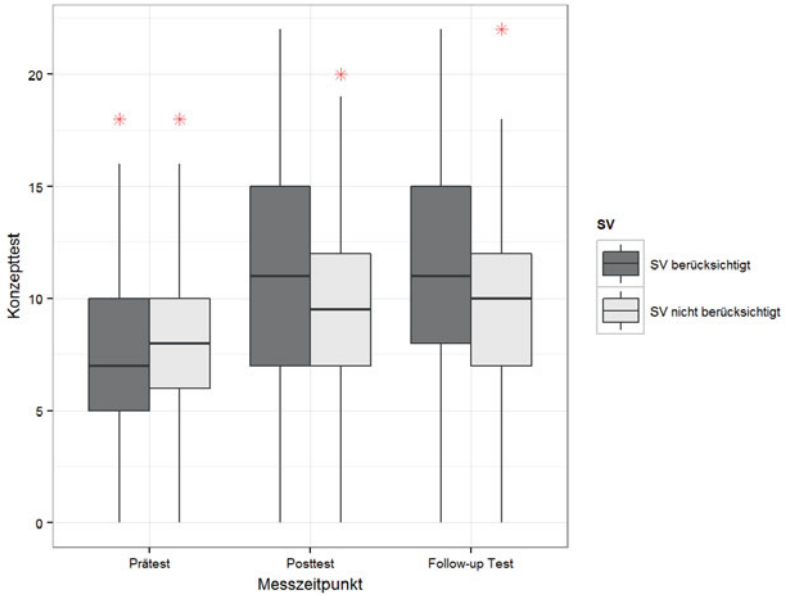


Abbildung 60: Boxplots zum grafischen Vergleich des konzeptuellen Verständnisses der beiden Stichproben je Messzeitpunkt

Auch der Verlaufsplot (vgl. Abbildung 61) veranschaulicht, dass in beiden Bedingungen das konzeptuelle Verständnis steigt und sich auf einem höheren Niveau stabilisiert, die Mittelwerte der Stichprobe „SV“ liegen dabei zu den späteren Messzeitpunkten über den Werten der Stichprobe „Ko“ (SV nicht brücksichtigt). Während im unteren Bereich (unterhalb des Mittelwertes, hier angezeigt durch die Linien) Schüler beider Stichproben vertreten sind, scheinen in den oberen Bereichen Schüler der Stichprobe „SV“ zu überwiegen. Daher empfiehlt es sich ATI-Effekte zu untersuchen.

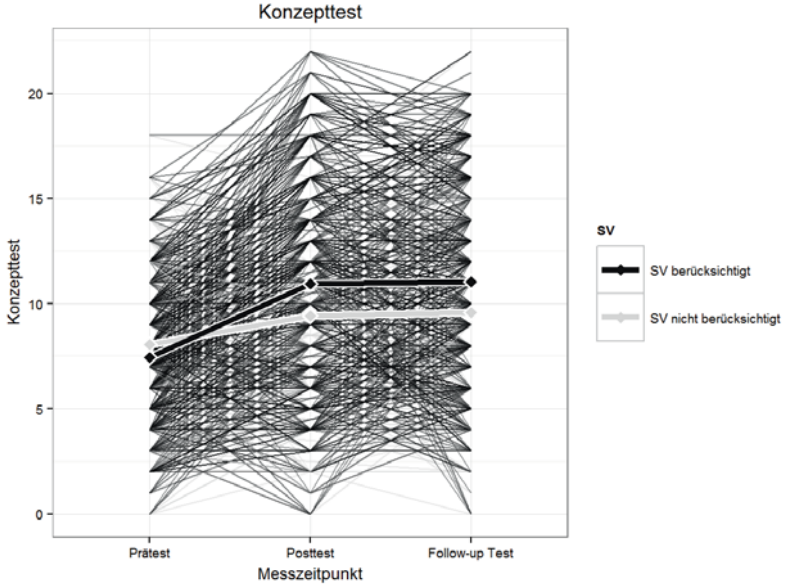


Abbildung 61: Verlaufsplot zum grafischen Vergleich des konzeptuellen Verständnisses der beiden Stichproben je Messzeitpunkt

Zu prüfen ist, ob diese Unterschiede im Lernzuwachs prä – post bzw. prä – follow-up signifikant sind: Analog zu den Mehrebenenmodellen zur Physikleistung und zum konzeptuellen Verständnis der Gesamtstichprobe „SV“ wurde ein Modell aufgestellt, das sowohl die Klassenstruktur, die Messwiederholung durch die Individuenkomponente als auch die Heteroskedastizität berücksichtigt und eine unstrukturierte Korrelationsmatrix der Fehlerterme im Individuum zugrunde legt.

Folgende Kovariaten fanden Eingang in das Modell, da hierzu in beiden Studien die gleichen Werte zur Verfügung standen: Schulnoten (letzte Zeugnisnote) in den Fächern Mathematik, Deutsch und Physik, Geschlecht und Klassengröße. Die Kovariaten kognitive Fähigkeiten (Intelligenzsubskalen) und Lehrerengagement aus Schülersicht (LES) entfielen wegen mangelnder Vergleichbarkeit in der Erhebung. Da an der Studie von Scheid (2013) keine Gesamtschüler teilnahmen, war es nicht sinnvoll die Variable Schultyp einzubeziehen.

Tabelle 61 Ergebnisse der Mehrebenenanalyse zum konzeptuellen Verständnis: Stichprobenvergleich „SV“ und „Ko“

Level (Stichprobengröße)	Erläuterung							
Level 1 *(N = 2692)*	Messzeitpunkte							
Level 2 *(N = 988)*	Individuen: Noten, Intelligenz, LES (= Lehrerengagement aus Schülersicht)							
Level 3 *(N = 37)*	Schulklassen: Bedingung Treatment- (TG) vs. Kontrollgruppe (KG), Schultyp, Klassengröße							
Fixe Effekte								
Variable	*Erläuterung*	*b*	*(SE)*	*β*	*(SE)*	$F_{(numDF,\ denDF)}$	*p*
Interzept	Durchschnittliches Verständnis: Stichprobe „Ko“, Geschlechtszugehörigkeit = weiblich, mit durchschnittlichen Noten	8.05	0.44	1.92	0.11	$320.82_{(1,1690)}$	< .001
Zuwachs (prä – post: t_2-t_1)	Durchschnittlicher Verständniszuwachs: Stichprobe „Ko“, Geschlechtszugehörigkeit = weiblich, mit durchschnittlichen Noten	1.11	0.26	0.26	0.06	$18.69_{(1,1690)}$	< .001
Zuwachs (prä – follow-up: t_3-t_1)	Durchschnittlicher Verständniszuwachs: Stichprobe „Ko“, Geschlechtszugehörigkeit = weiblich, mit durchschnittlichen Noten	1.79	0.24	0.42	0.06	$53.68_{(1,1690)}$	< .001
Bedingung = SV	Unterschied zwischen „SV“ und „Ko“ zu t_1	-0.90	0.61	-0.21	0.15	$2.19_{(1,34)}$	.148
Bedingung * Zuwachs prä – post	Zusätzlicher Einfluss Bedingung = „SV“ auf den Zuwachs zwischen t_2 und t_1	2.27	0.31	0.54	0.07	$51.94_{(1,1690)}$	< .001
Bedingung * Zuwachs prä – follow-up	Zusätzlicher Einfluss der Bedingung = „SV“ auf den Zuwachs zwischen t_3 und t_1	1.99	0.30	0.47	0.07	$45.10_{(1,1690)}$	< .001
Klassengröße	Genereller Einfluss auf Basis der Schätzung zu t_1	0.01	0.05	0.00	0.01	$0.15_{(1,34)}$	.907
Klassengröße *Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.01	0.02	-0.01	0.01	$0.30_{(1,1690)}$	.699
Klassengröße *Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.01	0.02	-0.01	0.01	$0.66_{(1,1690)}$	.585
Geschlecht * Ausgangswert	Genereller Einfluss Geschlechtszugehörigkeit = männlich auf Basis der Schätzung zu t_1	0.19	0.23	0.04	0.06	$0.36_{(1,947)}$	.417
Geschlecht * Posttest	Zusätzlicher Einfluss Geschlechtszugehörigkeit = männlich auf den Zuwachs zwischen t_2 und t_1	0.17	0.27	0.04	0.07	$0.30_{(1,1690)}$	.551
Geschlecht * Follow-up Test	Zusätzlicher Einfluss Geschlechtszugehörigkeit = männlich auf den Zuwachs zwischen t_3 und t_1	-0.35	0.26	-0.08	0.06	$1.85_{(1,1690)}$	.174
Mathematiknote	Genereller Einfluss auf Basis der Schätzung zu t_1	-0.31	0.13	-0.07	0.03	$6.19_{(1,947)}$	.013

Mathe-Note * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.10	0.18	-0.02	0.04	$0.31_{(1,1690)}$	.575
Mathe-Note * Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.39	0.17	-0.09	0.04	$5.42_{(1,1690)}$	.020
Deutschnote (D, M, PH)	Genereller Einfluss auf Basis der Schätzung zu t_1	-0.13	0.13	-0.03	0.03	$0.87_{(1,947)}$	.351
Deutschnote * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.17	0.19	-0.04	0.04	$0.78_{(1,1690)}$	.376
Deutschnote* Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.12	0.18	-0.24	0.04	$0.45_{(1,1690)}$	.501
Physiknote (D, M, PH)	Genereller Einfluss auf Basis der Schätzung zu t_1	-0.11	0.12	-0.03	0.03	$1.00_{(1,947)}$	.319
Physiknote * Posttest	Zusätzlicher Einfluss auf den Zuwachs zwischen t_2 und t_1	-0.83	0.16	-0.21	0.04	$26.23_{(1,1690)}$	< .001
Physiknote * Follow-up Test	Zusätzlicher Einfluss auf den Zuwachs zwischen t_3 und t_1	-0.51	0.14	-0.13	0.04	$11.95_{(1,1690)}$	< .001

Zufällige Effekte			
$\sigma_{u\ (Klasse)}$	1.50		
$\sigma_{p\ (Individuum)}$	1.70		
$\sigma_{e\ (individuenspezifisch)}$	2.27		
ρ *(Messzeitpunkt)*	(ρ_{12}) -0.05	(ρ_{13}) -0.03	(ρ_{23}) 0.27
g *(Messzeitpunkt)*	(g_1) 1.00	(g_2) 1.40	(g_3) 1.36

Modellvergleich	*Berichtetes Modell*	*Leermodell*[a]
Devianz	13883.94	14136.72
df	29	11
AIC	13941.94	14158.72
BIC	14112.99	14223.60
LR-Test *(berichtetes Modell versus Leermodell)*	45.89, p < .001	
Unerklärte Varianz Level 1	(t_1) 10.30 (t_2) 15.23 (t_3)14.70	(t_1) 11.27 (t_2) 17.63 (t_3) 16.95
Unerklärte Varianz Level 2	13.41	15.29
Unerklärte Varianz Level 3	2.69	3.24

Erklärte Varianz des berichteten Modells im Vergleich zum Leermodell

R^2 Level 1	(t_1) 0.09	(t_2) 0.14	(t_3) 0.13		
R^2 Level 2	0.12				
R^2 Level 3	0.17				
„Wirkung" des Treatments	Anteil aufgeklärter Varianz[b]			Effektstärke	
R^2 Level 1	(t_1) 0.01	(t_2) 0.02	(t_3) 0.01	$\Delta_{(prä-post)} = 0.63$	$\Delta_{(prä-follow\text{-}up)} = 0.56$
R^2 Level 2	0.01				
R^2 Level 3	< 0.01				

[a] Das Leermodell unterscheidet sich vom berichteten Modell darin, dass alle fixen Effekte außer dem Interzept und den Dummy-Variablen $t_2$und t_3 fehlen. Das Modell sagt für jeden Schüler voraus, dass seine Leistung zum jeweiligen Messzeitpunkt dem mittleren konzeptuellen Verständnis der Stichprobe entspricht - unter Berücksichtigung der individuen-und klassenspezifischen Unterschiede. Man beachte, das Leermodell schätzt folgende 11 Parameter (df = 11): Interzept, t_2, t_3, g_2, g_3, ρ_{12}, ρ_{13}, ρ_{23}, σ_u, σ_p, σ_ε.
[b] Der Anteil aufgeklärter Varianz errechnet sich wie in Tabelle 52 angegeben nach Snijder und Bosker (1999, S. 99 ff.).
Für das harmonische Mittel der Klassengrößen ergibt sich (h = 26.47).

Die Ergebnisse der Mehrebenenanalyse zeigen (vgl. Tabelle 61), dass die auf Basis der deskriptiven Statistiken zu vermutenden Mittelwertsunterschiede bei den Startbedingungen zwischen den Stichproben „SV“ und „Ko“ nicht signifikant sind ($F_{(1,34)} = 2.19$, n.s.). Die Null-Hypothese, dass Schüler in beiden Bedingungen mit dem gleichen Ausgangsniveau starten, wird beibehalten.

Der Vergleich des Lernzuwachses ergibt, dass es im weiteren Entwicklungsverlauf zu höchst signifikanten Unterschieden kommt. So erreichten Schüler, welche Unterricht besucht hatten, der gezielt auf Schülervorstellungen einging (Stichprobe „SV“), einen Lernzuwachs prä – post, der durchschnittlich um $b = 2.06$ Punkte höher lag und einen Lernzuwachs prä – follow-up, der um $b = 2.27$ Punkte den Lernzuwachs der Schüler in der Stichprobe „Ko“ übertraf. Dies entspricht einem Lernzuwachs prä – post ($\beta_{post} = 0.54$), der in etwa eine halbe Standardabweichung über dem Lernzuwachs der Stichprobe „Ko“ ($F_{(1,1690)} = 51.94$, $p < .001$) lag und einem Zuwachs prä – follow-up, der ebenfalls knapp um eine halbe Standardabweichung ($\beta_{follow\text{-}up} = 0.47$) den Zuwachs der Stichprobe „Ko“ ($F_{(1,1690)} = 45.10$, $p < .001$) übertraf. Es ergibt sich nach Bortz und Döring (2005) ein relevanter mittlerer Effekt von $\Delta_{(prä - post)} = 0.63$ und $\Delta_{(prä - follow\text{-}up)} = 0.56$.

Von den analysierten Kovariaten erwiesen sich die Mathematik- und die Physiknote als signifikante Variablen zur Erklärung des Abschneidens im Konzepttest. Schüler, deren Mathematiknote um eine volle Notenstufe besser liegt, erzielten über alle Messzeitpunkte hinweg ein Testergebnis, das um einen drittel Punkt höher lag ($b = -0.31, \beta = -0.07, F_{(1, 947)} = 6.19, p < .05$) als das Ergebnis ihrer Altersgenossen. Schüler, deren Mathematiknote um eine Notenstufe besser ist, erzielten darüber hinaus im Schnitt noch zusätzlich einen Lernzuwachs prä – follow-up, der ebenfalls um $b = 0.39$ Punkte höher lag, was in etwa einem Zehntel einer Standardabweichung entspricht ($\beta = -0.09$) als das Ergebnis ihrer Altersgenossen ($F_{(1,1690)} = 5.42, p < .05$). Schüler, deren Physiknote um eine volle Notenstufe besser liegt, erzielten im Schnitt zusätzlich einen Lernzuwachs prä – post, der um knapp einen Punkte höher liegt ($b = -0.83, \beta = -0.21, F_{(1,1690)} = 26.23, p < .001$) und einen Lernzuwachs prä – follow-up, der einen halben Punkt höher liegt als das Ergebnis ihrer Altersgenossen ($b = -0.51, \beta = -0.13, F_{(1,1690)} = 11.95, p < .01$). Die Leistungen im Fach Deutsch, die Klassengröße und die Geschlechtszugehörigkeit spielen gemäß der Ergebnisse des Modells (vgl. Tabelle 61) keine signifikante Rolle.

2.3.9.8 Vertiefende Analysen zu Aptitude-Treatment-Interaktionen

Abschließend wurde untersucht, ob die gewählte Lehr-Lernmethode (Treatment- versus Kontrollbedingung) mit ausgewählten Schülermerkmalen interagierte. Solche Wechselwirkungs-Effekte zwischen Begabungen bzw. Fähigkeiten und Lehrmethoden werden in der pädagogischen Psychologie unter dem Fachbegriff der Aptitude-Treatment-Interaktion diskutiert (vgl. Cronbach & Snow, 1977; Hasebrook, 2006).

Insbesondere wurde der Frage nachgegangen, ob Schüler mit höheren kognitiven Fähigkeiten oder besseren Noten vom Unterricht der Treatmentbedingung in stärkerem Ausmaß profitierten. Neben möglichen ATI-Effekten wurde zusätzlich auch die Interaktion zwischen Schultyp und Bedingung sowie Geschlechtszugehörigkeit und Bedingung untersucht. Hierzu wurde ein Mehrebenenmodell aufgestellt, welches Interaktionsterme zwischen Bedingung und dem jeweiligen Schülermerkmal zu den verschiedenen Messzeitpunkten enthielt.

Zur Analyse möglicher ATI-Effekt zwischen Schülermerkmalen und Treatment bezüglich der Physikleistung bei repräsentationalen Aufgaben und des konzeptuellen Verständnisses wurde folgenden Fragen nachgegangen:

- Profitieren Schüler mit besseren Schulleistungen in höherem Maß vom Treatment als Schüler mit schlechteren Schulleistungen?
- Profitieren Schüler mit besseren Testergebnissen im verbalen, im figural-räumlichen oder im figural-logischen Schlussfolgern deutlicher vom Treatment als Schüler mit geringeren Testergebnissen in diesen Bereichen?

Weitere Wechselwirkungen:

- Profitieren Gymnasiasten deutlicher von der Treatmentbedingung als Gesamtschüler?
- Profitieren Jungen in höherem Maß vom Treatment als Mädchen oder umgekehrt?

Auf Basis der Ergebnisse zu den Kovariaten des Mehrebenenmodells zur Physikleistung könnten Interaktionseffekte zwischen Schulleistung und Treatment sowie Schultyp und Treatment bestehen. So erwiesen sich Schulleistungen und Schultyp als höchst signifikante Merkmale zur Vorhersage der Physikleistung und des Lernzuwachses.

Für die verwendeten Subskalen des I-S-T 2000 R ist ein Interaktionseffekt am wahrscheinlichsten zwischen Treatment und figural-räumlicher Intelligenz zu erwarten, da sich das figural-räumliche Schlussfolgern signifikant auf den Lernzuwachs prä – post auswirkte. Eine Interaktion zwischen verbalem respektive figural-logischen Schlussfolgern und Treatment ist weniger wahrscheinlich, da das verbale Schussfolgern keinen signifikanten Erklärungswert für den Lernzuwachs prä – post bzw. prä – follow-up aufwies und sich für das figural-logische Schlussfolgern über alle Messzeitpunkte hinweg kein signifikanter Erklärungswert im Gesamtmodell zeigte.

Die Geschlechtszugehörigkeit erwies sich in den bisherigen Ergebnissen (siehe Kapitel 2.3.9.3 Erste Hypothese: Wirkung des Treatments auf die Physikleistung) nicht als signifikante Variable zur Vorhersage der Physikleistung. So schnitten Jungen zwar etwas besser ab als Mädchen und erzielten auch leicht höhere Lernzuwächse, diese Effekte waren aber jeweils nicht signifikant. Um der Vollständigkeit willen wurde jedoch auch die Interaktion zwischen Geschlechtszugehörigkeit und Treatment geprüft.

Im Ergebnis (siehe Tabelle 22 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com) konnte lediglich zwischen vorherigen Schulleistungen (Noten) und Bedingung ein signifikanter Interaktionseffekt bezüglich des Lernzuwachses festgestellt werden. „Gute Schüler“ (d.h. Schüler mit besseren Noten) in der Treatmentbedingung erzielten einen höheren Lernzuwachs prä – follow-up als „gute Schüler“ in der Kontrollbedingung (b = -1.45, β = -0.22, $F_{(1,800)}$ = 10.03, $p < .01$). Für den Lernzuwachs prä – follow-up bestand jedoch kein ATI-Effekt. Zwischen den übrigen Merkmalen: Intelligenz, Schultyp, Geschlechtszugehörigkeit und Treatment waren zu keinem Zeitpunkt ATI-Effekte nachweisbar.

Im Hinblick auf die Entwicklung des konzeptuellen Verständnisses wurden ebenso wie für die Entwicklung der Physikleistung bei repräsentationsbezogenen Aufgaben die oben genannten Fragen untersucht.

Im Ergebnis (siehe Tabelle 23 Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com) wurde lediglich zwischen Geschlechtszugehörigkeit und Bedingung ein signifikanter Interaktionseffekt bezüglich des Lernzuwachses prä – follow-up festgestellt. So erzielten Mädchen in der Treatmentbedingung einen signifikant höheren Lernzuwachs im konzeptuellen Verständnis prä – follow-up als Jungen in der Kontrollbedingung (b = -1.94, β = -0.42, $F_{(1,788}$= 5.77, $p < .05$). Da Jungen jedoch über alle Messzeitpunkte hinweg einen etwas höheren (wenn auch nicht signifikant höheren) Lernzuwachs prä – follow-up erreichten (b = 0.76, β = 0.16, $F_{(1,788)}$ = 1.63, *n.s.*), kann nicht die Schlussfolgerung gezogen werden, dass Mädchen in der Treatmentbedingung in der Summe von einem signifikant

höheren Lernzuwachs profitierten. Die Interaktion zwischen Geschlechtszugehörigkeit und Bedingung zeigte sich auch nicht für den Lernzuwachs prä – post. Zwischen den übrigen Merkmalen: Intelligenz, Schulnoten und Schultyp waren zu keinem Zeitpunkt ATI-Effekte nachweisbar.

Auch für die abhängige Variable Schülermotivation wurden mögliche Interaktions-Effekte untersucht. Im Gesamtmodell hatten sich die Fachnoten, die Geschlechtszugehörigkeit, der Schultyp und das Lehrerengagement aus Schülersicht (LES) als relevante Kovariaten erwiesen. Auf Basis dieser Ergebnisse wurde folgenden Fragen nachgegangen:

- Unterscheidet sich die motivationale Entwicklung im Schuljahresverlauf von Gymnasiasten in der Treatmentbedingung von der Entwicklung der Gesamtschüler in der Treatmentbedingung?
- Steigern Schüler mit besseren Schulleistungen (guten Noten) in der Treatmentbedingung ihre Motivation deutlicher als Schüler mit besseren Schulleistungen (guten Noten) in der Kontrollbedingung?
- Unterscheidet sich die motivationale Entwicklung im Schuljahresverlauf von Schülern mit besseren Testergebnissen im figural-logischen Schlussfolgern in der Treatmentgruppe von Schüler mit geringeren Testergebnissen in diesem Bereich?
- Unterscheidet sich die motivationale Entwicklung im Schuljahresverlauf von Jungen in der Treatmentbedingung von der Entwicklung der Mädchen in der Treatmentbedingung?
- Unterscheidet sich die motivationale Entwicklung im Schuljahresverlauf von Schülern in der Treatmentbedingung, die ihren Lehrer als engagierter wahrnehmen, von Schülern in der Kontrollbedingung?

Im Ergebnis (siehe Tabelle 24 Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com) konnte lediglich zwischen LES und Bedingung ein signifikanter Interaktionseffekt bezüglich des Motivationsverlaufs prä – follow-up festgestellt werden. Die Motivation von Schülern in der Treatmentbedingung, die ihren Lehrer als engagierter wahrnehmen, sank vom Prä- zum Follow-up Test stärker als die Motivation von Schülern in der Kontrollbedingung (b = -0.14, β = -0.21, $F_{(1,806)} = 7.96$, $p < .01$).

Interessanterweise ergab sich kein genereller Interaktionseffekt (über alle Messzeitpunkte hinweg) zwischen Bedingung und LES. Auch für die Entwicklung prä – post ist kein Interaktionseffekt nachweisbar. Eine Interaktion zwischen Treatment und LES ist also im Anschluss an die Unterrichtsreihe nicht zu beobachten.

Offenbar werden jedoch Schüler in der Treatmentbedingung, die ihren Lehrer als motivierter wahrnehmen, mittelfristig (prä – follow-up) leicht demotiviert: Liegt die anfänglich wahrgenommene Einschätzung des Lehrerengagements eine Standardabweichung über dem Mittelwert der Gesamtstichprobe, kommt es zu einem mittelfristig nachweisbaren Motivationsverlust der einem Fünftel einer Standardabweichung entspricht. Dies drückt sich in einem Verlust von b = -0.14 Punkten auf einer Skala von 1 bis 6 aus.

Zwischen den übrigen Merkmalen: Schulleistungen (Fachnoten), Schultyp, figural-logische Intelligenz, Geschlechtszugehörigkeit und Treatment waren zu keinem Zeitpunkt Interaktions-Effekte in Bezug auf die Motivation im Schuljahresverlauf nachweisbar.

Abschließend wurde auch für den Stichprobenvergleich der Frage nachgegangen, ob zwischen Fachleistungen bzw. Begabungen eine Aptitude-Treatment-Interaktion vorliegen könnte. Da in den beiden Studien der Intelligenztest nicht identisch war, wurden die jeweiligen Fachnoten in Mathematik, Deutsch und Physik herangezogen und für jede einzelne Fachnote der jeweilige Interaktionsterm überprüft. Im Ergebnis (siehe Tabelle 25 in Anhang C9 auf der Produktseite zu diesem Buch unter www.springer.com) war jedoch zu keinem der Messzeitpunkt ein ATI-Effekt nachweisbar.

2.3.9.9 Zusammenfassung der Ergebnisse zur Untersuchung der Hypothesen 1-5

Abbildung 62 veranschaulicht die Ergebnisse zu den Hypothesen 1-4. Sie fasst alle untersuchten signifikanten Einflussfaktoren auf die drei abhängigen Variablen zusammen. Die Darstellung ist nicht mit einem Strukturgleichungsmodell zu verwechseln. Die verwendeten Pfeile beziehen sich auf diejenigen standardiserten Regressionsgewichte (β), welche signifikante Prädiktoren für das jeweillige Kriterium darstellen: (1) Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen, erfasst im Leistungstest (LT), (2) das konzeptuelle Verständnis, erfasst im Konzepttest (KT) und (3) die Schülermotivation, erhoben im Motivationsfragebogen (Mo).

Wie sich an Abbildung 62 ablesen lässt, ist die Bedingung (Treatment- versus Kontrollgruppe) lediglich für die Teilstichprobe der Gymnasiasten (hier dargestellt durch einen grauen Pfeil) ein signifikanter jedoch eher schwacher Prädiktor mit einer im Vergleich zu den übrigen Prädiktoren eher niedrigen Ausprägung von $\beta = 0.16, p < .05$.

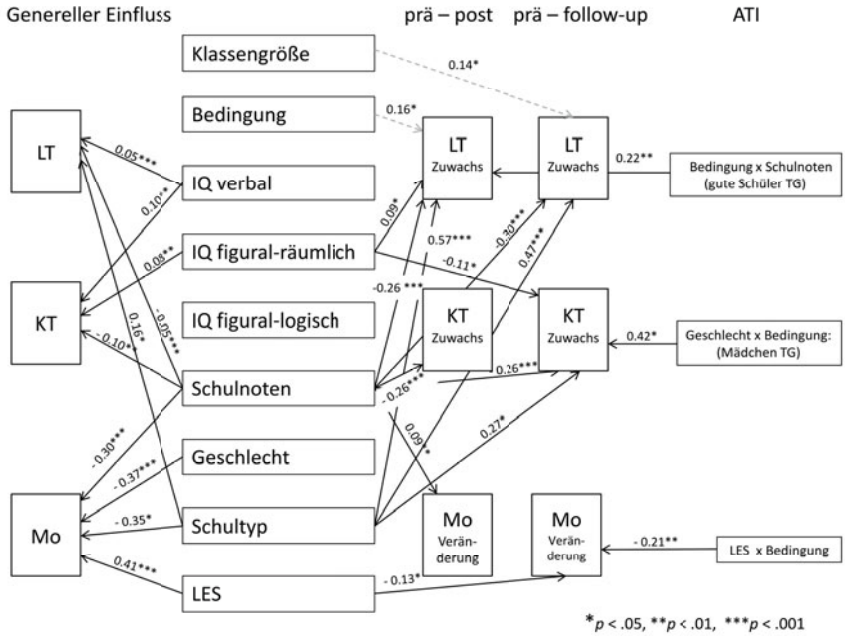


Abbildung 62: Zusammenschau relevanter Prädiktoren und Aptitude-Treatment-Interaktionen (ATI) in Bezug auf die Test- bzw. Fragebogenergebnisse im Physikleistungstest (LT), Kontzepttest (KT) und die Schülermotivation (Mo). Gestrichelte Pfeile beziehen sich auf Prädiktoren, die lediglich in der Teilstichprobe der Gymnasiasten relevant waren.

Von den übrigen Variablen sind für die Vorhersage der leistungsbezogenen abhängigen Variablen Physikleistung (LT) und konzeptuelles Verständnis (KT) vor allem die Fachnoten aussagekräftig. Verbales und figural-räumliches Schlussfolgern sowie der Schultyp spielen ebenfalls eine Rolle. In Bezug auf die Motivation sind die Geschlechtszugehörigkeit, die Fachnoten und das Lehrerengagement aus Schülersicht relevant. Klassengröße und figural-logisches Schlussfolgern spielen eine untergeordnete oder überhaupt keine Rolle.

Die Ergebnisse zum Vergleich der inhaltlich aufeinander bezogenen Studien „SV“ versus „Ko“ (Scheid, 2013) – Hypothese 5 – sind in Abbildung 63 zusammenfassend dargestellt.

Es zeigt sich, dass die Bedingung (wurden Schülervorstellungen thematisiert oder nicht) in vergleichsweise hohem Maß entscheidend für die Erklärung des Zuwachses des konzeptuellen Verständnisses ist.

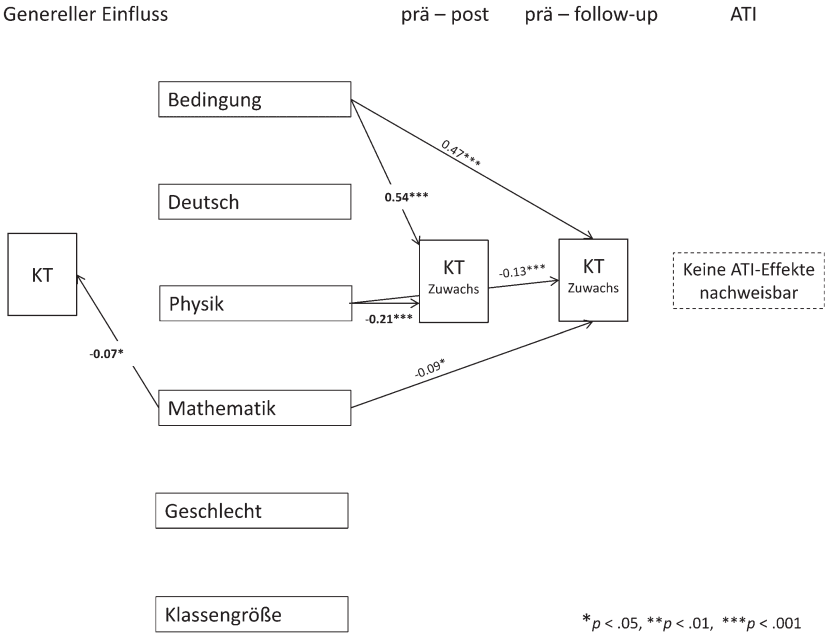


Abbildung 63: Zusammenschau relevanter Prädiktoren für den Vergleich inhaltlich aufeinander bezogener Studien „SV“ versus „Ko“ in Bezug auf die Test- bzw. Fragebogenergebnisse im, Konzepttest (KT)

Von den Kovariaten ist (erwartungsgemäß) die Physiknote ein aussagekräftiger Prädiktor, die Mathematikvorleistung (ebenfalls erfasst durch die letzte Zeugnisnote) spielt ebenfalls eine signifikante jedoch weniger ausgeprägte Rolle. Die Deutschvorleistung, die Klassengröße und Geschlechtszugehörigkeit sind zu keinem Zeitpunkt signifikante Einflussfaktoren für das Abschneiden im Konzepttest.

2.3.9.10 Diskussion der Ergebnisse zur Untersuchung der Hypothesen 1-5

Im Rahmen der Hauptstudie wurden fünf Hypothesen untersucht, welche sich auf die drei abhängigen Variablen beziehen: Wissen und Problemlösen bei repräsentationsbezogenen Aufgaben, das konzeptuelle Verständnis in der Strahlenoptik und die Motivation der Schüler im Physikunterricht.

Hypothese 1: Die Treatmentgruppe erreicht im Wissen und Problemlösen beim Umgang mit fachspezifischen Repräsentationen im Bereich der Strahlenoptik unmittelbar nach der Intervention und ca. zwei Monate später einen höheren Lernzuwachs als die Kontrollgruppe.

In dem Mehrebenenmodell zur Physikleistung bei repräsentationsbezogenen Aufgaben zeigte sich kein signifikanter Unterschied zwischen Treatment- und Kontrollgruppe für die Gesamtstichprobe. Die Verlaufsdiagramme veranschaulichen, dass sich Wissen und Problemlösen in beiden Gruppen parallel entwickelten. Die erste Hpothese kann daher nicht bestätigt werden

In Bezug auf die Kovariaten fielen die heterogenen Ergebnisse bezüglich der kognitiven Fähigkeiten auf, so erwies sich die verbale Intelligenz über alle Messzeitpunkte hinweg als relevant, wirkt sich aber nicht auf den Lernzuwachs aus. Die figural-räumliche Intelligenz beeinflusst signifikant nur den Lernzuwachs prä – post. Beides lässt sich plausibel erklären, so mag ein gutes Verständnis des Aufgabentextes Schüler prinzipiell bei der Bearbeitung unterstützen. Für die Erklärung des Lernzuwachses zu den beiden späteren Messzeitpunkten ist diese Fähigkeit wenig relevant. Hier entscheidet vermutlich ein gutes räumliches Vorstellungsvermögen darüber, ob fachlich adäquate mentale Modell gebildet werden.

Der Einfluss der Klassengröße bei Gymnasiasten auf den Lernzuwachs prä – follow-up wird als Artefakt gewertet (Schüler in größeren Klassen schneiden besser ab). Dafür spricht, dass weder ein Einfluss der Klassengröße über alle Messzeitpunkte hinweg noch ein signifikanter Einfluss der Klassengröße auf den Lernzuwachs prä – post festgestellt werden konnte. Jedoch erzielen Schüler in größeren Klassen an Gymnasien einen höheren Lernzuwachs prä – follow-up als Schüler in kleineren Klassen. Vermutlich kovariieren hier Einflussfaktoren auf Klassen- und Lehrerebene systematisch mit Variablen, die dafür sorgen, dass Schüler aus größeren Klassen weniger Lerninhalte vergessen als Schüler aus kleineren Klassen, z.B. könnte es in „besseren“ Gymnasien mehr Schüler-Anmeldungen gegeben haben. In allen übrigen Modellen spielt die Klassengröße keine Rolle. Hier bestätigt sich das aus der Bildungsforschung bekannte Ergebnis, dass kleine Klassen für den Lernerfolg nicht von signifikantem Vorteil sind (vgl. auch Ergebnisse der dritten PISA-Studie 2006, PISA-Konsortium Deutschland, 2007).

Unterschiedliche Ergebnisse zeigten sich für die Teilstichprobe der Gymnasiasten, die zu allen Messzeitpunkten im Schnitt bessere Ergebnisse erzielten als die Gesamtschüler. Hier ergaben sich unmittelbar nach der Intervention signifikante Vorteile zu Gunsten der Treatmentgruppe, welche jedoch mittelfristig nicht bestehen blieben. Die resultierende Effektstärke war jedoch gering, so dass

dieser Befund für sich genommen für die Unterrichtspraxis zunächst wenig bedeutsam erscheint.

Insgesamt wird die erste Hypothese – mit Ausnahme des Befundes der signifikanten Vorteile der Treatmentgruppe an Gymnasien – nicht unterstützt. Entweder spielt das Ausmaß an kognitiver Aktivierung keine Rolle für die untersuchten Lernprozesse oder die Unterschiede im Ausmaß der kognitiven Aktivierung waren nicht klar genug operationalisiert. Die zuletzt genannte These wird von den Einschätzungen der Lehrkräfte (Zustimmung zur Frage: „Ich schätze, der Unterricht hat die Schüler zum Nachdenken angeregt“) gestützt. So zeigten sich im Mittel für beide Gruppen gleich hohe Werte in der Zustimmung zu dem Item, obgleich sich geringere Maße der Übereinstimmung für die Bewertung des Unterrichts in der Kontrollgruppe ergaben. Wäre der Unterschied deutlicher zum Tragen gekommen, hätte die Zustimmung zu dem Item in der Treatmentgruppe signifikant höher ausfallen müssen als in der Kontrollgruppe.

Hypothese 2: Die Treatmentgruppe erreicht im konzeptuellen Verständnis für den Bereich der Strahlenoptik unmittelbar nach der Intervention und ca. zwei Monate später einen höheren Lernzuwachs als die Kontrollgruppe

In den Mehrebenenmodellen zur Analyse des konzeptuellen Verständnisses zeigten sich ebenfalls keine signifikanten Unterschiede zwischen den Bedingungen TG und KG.

Im Gegensatz zum Physikleistungstest ergaben sich zwischen dem Schultyp Gymnasium versus IGS kein bzw. ein geringer Vorteil der Gymnasiasten für den Vergleich prä – follow-up. Dieser Befund kann als Hinweis gewertet werden, dass das Abschneiden im Konzepttest weniger von der generellen Leistungsfähigkeit beeinflusst wird, welche bei den Gymnasien durchschnittlich höher ausgeprägt sein dürfte: Mehrere Indikatoren (höhere Werte in den drei Subskalen des I-S-T 2000 R und generell bessere Fachnoten) weisen auf eine höhere Leistungsfähigkeit der Gymnasiasten hin.

Im Resümee konnte die zweite Hypothese nicht bestätigt werden. Da in beiden Bedingungen Schülervorstellungen thematisiert wurden, war vermutlich hier ebenso wie für den Leistungstest das Ausmaß an kognitiver Aktivierung entweder wenig entscheidend oder nicht deutlich genug operationalisiert.

Hypothese 3: Ein Teil der Treatmentgruppe, die ein variiertes Treatment erhielt, welches die Predict-Observe-Explain-Strategie (POE-Sequenz) unter der Perspektive des Lernens mit multiplen Repräsentationen beinhaltet, erreicht im Vergleich zu dem regulären Treatment und im Vergleich zur Kontrollgruppe einen höheren Lernzuwachs im konzeptuellen Verständnis unmittelbar nach der Intervention und ca. zwei Monate später.

Ein positiver Effekt des Einsatzes der POE-Sequenz auf den Lernerfolg der Schüler wurde nur sehr eingeschränkt nachgewiesen (Analyse der dritten Hypothese: Vergleich des regulären mit dem variierten Treatment), obgleich etliche andere Studien deutliche signifikante Effekte berichten (vgl. White & Gunstone, 1992; Kearney et al., 2001; Crouch et al., 2004 u.a.).

Möglicherweise war jedoch die relative Lernzeit innerhalb der Unterrichtssequenz zu gering, da die jeweiligen POE-Sequenzen nur in zwei von sechs Stunden und auch nur innerhalb eines Teils der Schulstunden zum Einsatz kamen.

Hypothese 4: Die Entwicklung der Motivation der Schüler in der Treatmentgruppe unterscheidet sich zu keinem der drei Messzeitpunkte von der Entwicklung in der Kontrollgruppe.

In Bezug auf die Entwicklung der Motivation waren keine Unterschiede feststellbar (Ablehnung der Unterschiedshypothese). Die intendierte geringere kognitive Aktivierung in der Kontrollgruppe wirkt sich somit nicht negativ auf die Motivation der Schüler aus. Für die Treatmentgruppe ergibt sich umgekehrt auch keine Verbesserung der Motivation (prä – post). Die generelle Abnahme der Motivation im Schuljahresverlauf wird zwar als bedauerlich, jedoch auch als wenig problematisch bewertet, da sie etlichen Ergebnissen zum Motivationsverlust im Schuljahresverlauf und in der weiteren Schulkarriere entspricht, welche aus der Bildungsforschung bekannt sind (vgl. Hidi, 2002; Wild & Hofer, 2000; Krapp, 2002; zitiert nach Schiefele, 2009, S. 171).

Hypothese 5: Schüler, die ein Treatment zur Förderung des konzeptuellen Verständnisses erhielten (TG und KG dieser Studie), erzielten einen höheren Lernzuwachs in Bezug auf das konzeptuelle Verständnis als Schüler, die ein Treatment zur Förderung der Kohärenz (Scheid, 2013) erhielten bzw. der Kontrollgruppe der entsprechenden Stichprobe angehörten.

Der Vergleich der beiden Studien innerhalb des Projekts zeigt, dass repräsentationsbezogene Aufgaben, welche Schülervorstellungen adressieren, zu einer signifikanten Verbesserung des konzeptuellen Verständnisses führen, die mittelfristig (zwei Monate später) erhalten bleibt (Bekräftigung der Ergebnisse der Pilotstudie). Die fünfte Hypothese wird bestätigt. Der resultierende Effekt wird als relevanter, mittlerer Effekt bewertet. Im Vergleich zu den Ergebnissen der Pilotstudie ergibt sich eine Relativierung des Befundes eines starken Effekts, der darauf zurückgeführt werden könnte, dass auch die Schüler in der Vergleichsstichprobe deutlicher von der Intervention profitieren, was sich in den Lernzuwächsen der Stichprobe „Ko“ zeigt.

Vertiefende Analysen zu ATI-Effekten: Obgleich die Verlaufsdiagramme das Bestehen von ATI-Effekten zwischen Treatment und Schultyp nahe legen, konnte

für keine der abhängigen Variablen ATI-Effekte gefunden werden. Offenbar wirkt das Treatment in beiden Schularten in vergleichbarer Weise, wobei sich das Leistungsniveau der Gesamtschüler im Mittel auf einem geringeren Niveau befinden. Bemerkenswerte ATI-Effekte bestanden lediglich zwischen Physikleistung und Schulnoten. Offenbar profitieren leistungsstärkere Schüler in höherem Maß von kognitiv aktivierenden Aufgaben im Umgang mit Repräsentationen als leistungsschwächere.

Eine mögliche Erklärung könnte darin bestehen, dass die aktive Auseinandersetzung mit fachspezifischen Repräsentationen ein bestimmtes Maß an Anstrengungsbereitschaft (und fachlichen Fähigkeiten) erfordert. Ist hier ein möglicher Schwellenwert überschritten, kommt es zur Bildung adäquater mentaler Modelle, welche zu einem höheren Niveau beim Wissen und Problemlösen führt.

Etwaige Effekte zwischen Schultyp und Schulnoten wurden vermutlich durch das höhere Niveau der Fachnoten an Gymnasien erklärt.

Die gefundene Interaktion zwischen LES und Treatment prä – post in Bezug auf die Motivation, sollte nicht überbewertet werden, da es sich um einen minimalen Unterschied handelt. Der positive Effekt für den Lernzuwachs des konzeptuellen Verständnisses der Mädchen, hebt sich, wie zuvor dargestellt, in der Summe auf.