Progress in Science Education
2021, Vol. 4 No. 1, 11-35
ISSN 2405-6057
DOI.10.25321/prise.2021.1015

KTSO-A: Konzepttest-Strahlenoptik – Abbildungen. Entwicklung eines Konzepttests zur Erfassung von Konzepten der Lichtausbreitung, Streuung und der Entstehung reeller Bilder im Bereich der Strahlenoptik

Rosa Hettmannsperger[1], Andreas Müller[2], Jochen Scheid[3], Jochen Kuhn[4] und Patrik Vogt[5]

[1]Hessische Lehrkräfteakademie, Wiesbaden
[2]Section Physique, Faculté des Sciences, Université de Genève
[3]Institut für naturwissenschaftliche Bildung, AG Physikdidaktik, Universität Koblenz-Landau, Campus Landau
[4]Fachbereich Physik/Didaktik der Physik, Technische Universität Kaiserslautern
[5]Institut für Lehrerfort- und -weiterbildung, Mainz
*Please address all correspondence to Rosa Hettmannsperger, rosa.hettmannsperger-lippolt@kultus.hessen.de

STRUCTURED ABSTRACT

Hintergrund: Die Erfassung von konzeptuellem Wissen und Verständnis ist ein Thema, das sowohl in zahlreichen physikdidaktischen Forschungsarbeiten als auch für die Diagnose des Kenntnisstands von Schülerinnen und Schülern im Unterrichtskontext relevant ist. Für den Bereich der Strahlenoptik liegt zwar seit Jahrzehnten eine breite Wissensbasis zu konzeptionellen Lernschwierigkeiten und Schülervorstellungen und darauf basierenden Items vor, sowie auch verschiedene Tests, die jedoch nur ansatzweise psychometrisch charakterisiert und validiert sind.

Ziel: Aus diesem Grund wurde im Rahmen einer Interventionsstudie im Physikunterricht der Sekundarstufe I ein Konzepttest zu den Teilbereichen Lichtausbreitung, Bildentstehung und Strahlenmodell/Bildkonstruktion im Kontext des Lerngegenstandes „Abbildung bei Sammellinse“ entwickelt und empirisch auf das Vorliegen akzeptabler Test-Charakteristika überprüft (Konzept-Test Strahlenoptik: Optische Abbildungen – KTSO-A).

Stichprobe: Das Testinstrument (bestehend aus 10 Items) wurde in zwei Teilstichproben von Schülerinnen und Schülern (n = 389 und n = 480) der Sekundarstufe I erprobt.

Methode: Die Testkonstruktion basiert auf Forschungsergebnissen empirischer Studien zu konzeptuellen Schwierigkeiten in dem betrachteten Teilbereich der Strahlenoptik und folgt der Literatur zum Vorgehen bei der Entwicklung von Konzepttests. Neben der Berechnung üblicher Test- und Itemkennwerte (Itemschwierigkeiten, Trennschärfen, Reliabilität) wurden im Rahmen einer Kreuzvalidierung zwischen den beiden Teilstichproben die Methoden der exploratorischen und konfirmatorische Faktorenanalysen kombiniert und zudem ein Expertenrating umgesetzt.

Ergebnisse: Die Kennwerte des Tests (Itemschwierigkeiten, Trennschärfen, Reliabilität) liegen für die beiden Stichproben durchwegs bzw. größtenteils in den in der Testtheorie üblichen Akzeptanzbereichen. Abweichungen davon, Unterschiede zwischen den Stichproben und Details auf Item- und Distraktorebene, lassen sich und unter Bezug auf die vorliegende Forschung zu Konzepttests sinnvoll interpretieren. Die inhaltliche und curriculare Validität des Tests wurde in einem Expertenrating als hoch eingestuft. Mittels des Verfahrens der Kreuzvalidierung wurde die Dimensionalität des Tests untersucht. Die Analyse zeigt eine dreidimensionale Struktur mit akzeptablen bis guten Kennwerten auf (gute Fit-Indizes: $\chi^2_{(32)}$ 44.13, p = 0.075; CFI = 0.98; TLI = 0.98, jedoch teilweise niedrige Faktorladungen) und ermöglicht eine plausible inhaltliche Deutung in folgenden Teildimensionen: Bildentstehung, geradlinige Lichtausbreitung und Streuung sowie Verständnis des Strahlenmodells und der Bildkonstruktion.

***Diskussion und Relevanz*:** Die Gesamteinschätzung des vorgestellten Tests wird im Hinblick auf den Stand der Entwicklung von Konzept-Tests in anderen Bereichen und dabei bekannte Begrenzungen und Schwierigkeiten bezüglich Validität, Reliabilität und Strukturaufklärung ausführlich diskutiert. Die psychometrischen Eigenschaften des KTSO-A können im Vergleich mit dem Entwicklungsstand von Konzepttests im Allgemeinen als akzeptabel bis gut eingeschätzt werden. Darüber hinaus werden einige mögliche Perspektiven für künftige Weiterentwicklungen diskutiert. Der Wert des vorgestellten Konzept-Tests sehen wir darin, dass er für einen inhaltlich und curricular wohldefinierten Bereich eine praktikable und nach gängiger Forschungspraxis im Feld psychometrisch adäquate Testfassung bereitstellt, die für Forschungszwecke und zur unterrichtlichen Diagnostik eingesetzt werden kann.

Keywords: *Konzepttest, Strahlenoptik, optische Abbildungen, Faktorenanalyse*

STRUCTURED ABSTRACT

Background: A number of several concept inventories have been developed for various domains of physics, such as mechanics, heat, electricity or astronomy. There can be obtained test instruments for the domain of ray optics, as well. However, these inventories were only partially psychometrically validated.

Purpose: For this reason, an inventory (KTSO-A Konzepttest Strahlenoptik-Abbildungen / Concept Inventory Ray Optics – Imaging), was created detecting conceptions in the domains of light propagation, image formation and understanding ray diagrams in the context of imaging via convex lens. An empirical investigation of item statistics and test parameters was carried out to check if all parameters were within acceptable range.

Sample: The developed inventory consisting of 10 items was investigated based on two samples of students in the 7th and 8th grade at secondary schools in Rhineland-Palatinate (n = 389 and n = 480).

Methods: The outlined instrument is based on a well-known body of results on students' difficulties with learning optics and related to intuitive conceptions reported for this domain. The method of test construction followed the procedure of creating inventories in physics education known in notable literature.

Results: An item analysis revealed that the inventory tasks allow us to gather students' knowledge and understanding with a satisfactory discriminatory power and reliability. In order to investigate the dimensionality of the test, exploratory and confirmatory factor analyses were carried out. Results ($\chi^2_{(32)}$ 44.13, p = 0.075; *CFI* = 0.98; *TLI* = 0.98, *RMSEA* = 0.05; *SRMR* = 0.03) indicate that there are three clearly interpretable subdimensions: understanding image formation, conceptions of linear light propagation and scattering as well as understanding ray diagrams.

Conclusion: In accordance with psychometric standards, the presented test version can be already used both in science education research and classrooms at schools. Finally, some prospects for future improvements of the presented inventory are outlined.

Keywords*: concept inventory, ray optics, imaging, factor analysis*

Received: *May 2020. Accepted: August 2020.*

1 EINLEITUNG

Forschungsergebnisse zum Thema Schülervorstellungen und Lernschwierigkeiten zeigen, dass das Denken vieler Schülerinnen und Schüler von inkonsistenten und fachlich unangemessenen Vorstellungen geprägt ist (Duit & Treagust, 2003; Duit, 2009; Scott, Adams & Leach, 2007; Shtulman & Lombrozo, 2016). Viele der fehlerbehafteten Vorstellungen der Schüler sind tief in den Alltagserfahrungen verankert, andere werden erst im Unterricht selbst gebildet. Klassische Lehrstrategien erweisen sich oft als wenig effektiv darin, die Vorstellungen der Lernenden zu verändern (Andersson & Kärrqvist, 1983; White & Gunstone, 1989; Fetherstonhaugh & Treagust, 1992; Galili, 1996; Langley, Ronen & Eylon, 1997; Limon, 2001; Heywood, 2005). Lernprozesse, die eine Wissensveränderung beinhalten, sind in der Regel zwar recht (zeit-)aufwändig, aber dennoch realisierbar. So belegen Guzzetti, Snyder, Glass und Gamas (1993) in einer Metaanalyse, dass Vermittlungsstrategien, in denen Konflikte zwischen den vorunterrichtlichen Vorstellungen und den zu lernenden wissenschaftlichen Vorstellungen herausgearbeitet worden sind, solchen Strategien statistisch signifikant überlegen waren, in denen das nicht der Fall war.

Die Feststellung, dass das konzeptuelle Verständnis von Lernenden verändert werden kann, ist jedoch daran geknüpft, dieses Verständnis erfassen zu können. Hier setzt die Entwicklung von Konzepttests an, wie sie für viele Gebiete der Naturwissenschaften bereits entwickelt wurden (Liu, 2012; Madsen, McKagan & Sayre, 2017). Auch im Bereich der (Strahlen-) Optik liegen eine Reihe von Wissens- und Verstehensfragen sowie einige Tests vor, die auf die Erfassung von konzeptuellem Wissen und Verständnis in der (Strahlen-) Optik zielen und hierbei weitverbreitete Schülervorstellungen und Lernschwierigkeiten einbeziehen (siehe Abschnitt 2.2). Bislang fehlen jedoch publizierte, psychometrische charakterisierte Testinstrumente, mit denen verschiedene Wissensfacetten im Bereich der Strahlenoptik in praktisch und methodologisch zufriedenstellender Weise geprüft werden können. Die vorliegende Arbeit soll einen Beitrag zur Schließung dieser Lücke zu den Teilbereichen Lichtausbreitung, Bildentstehung und Strahlenmodell/Bildkonstruktion leisten. Es wird zunächst ein Forschungsüberblick zu konzeptionellen Schwierigkeiten gegeben und dann die Entwicklung und psychometrische Charakterisierung eines einschlägigen Testinstrumentes vorgestellt.

2 FORSCHUNGSHINTERGRUND

2.1 Definitionen und Vorgehen

Konzeptuelles Wissen definiert sich nach (Byrnes & Wasik, 1991, S. 777) als Wissen über Kernkonzepte einer Domäne und die Beziehung dieser Kernkonzepte untereinander. Es zeichnet sich gemäß Byrnes und Wasik durch

die kompetente Verwendung unterschiedlicher Konstrukte aus, welche aus semantischen Netzwerken, Hierarchien und mentalen Modellen bestehen können. Baumert et al. (2001) betonen in Hinblick auf die PISA-Studien, dass Schülerinnen und Schüler konzeptuelles Verständnis benötigen, „um Phänomene der natürlichen und der von Menschen geschaffenen Welt zu verstehen“. Deshalb wird hier weniger Wert auf die Reproduktion von Faktenwissen gelegt, hingegen geht es darum, „ein konzeptuelles Verständnis“ zu erfassen, das mit der Anwendung von Alltagskonzepten beginnt und bis zum Arbeiten mit naturwissenschaftlichen Modellvorstellungen reicht.

Unter didaktischer Perspektive kann die Erfassung des Wissensstandes helfen, das Denken der Lernenden kennenzulernen, um den Unterricht entsprechend anpassen und abstimmen zu können (vgl. Tyson, Venville, Harrison & Treagust, 1997; Özdemir & Clark, 2007). Methoden, die einen Einblick in das Denken der Schülerinnen und Schüler und ihre Lernschwierigkeiten erlauben, wie beispielsweise offene Interviews oder halbstandardisierte Fragebögen, sind in der Forschungspraxis und im Schulalltag sowohl bei der Durchführung als auch bei der Auswertung mit einem hohen zeitlichen Aufwand verbunden. Tests im Multiple-Choice Format weisen den Vorteil auf, dass sie bei der Durchführung und Auswertung relativ wenig Zeit beanspruchen und dennoch einen Einblick in die Wissensstruktur der Lernenden ermöglichen, sofern sie entsprechend konzipiert sind und den Standards der Testkonstruktion entsprechen.

Üblicherweise bestehen Konzepttests aus Single- oder Multiple-Choice-Items, bei denen die Teilnehmer aus mehreren vorgegebenen Antwortmöglichkeiten diejenige(n) auswählen, die sie als zutreffend befinden (siehe Lindell, Peak & Foster, 2007; Liu, 2012 im Überblick). Neben der korrekten Antwortalternative erfordert die Entwicklung der Items also auch die Formulierung geeigneter, d. h. „plausibler“ Distraktoren. Diese Distraktoren basieren in der Regel auf weitverbreiteten fehlerhaften Schülervorstellungen (vgl. ebd.). Gemeint sind hierbei Vorstellungen, „die einem wissenschaftlichen Verständnis entgegenstehen oder so unklar und vage sind, dass sie einer fachlichen Präzisierung bedürfen“ (Hettmannsperger, 2015, S. 67).

Für einige Themengebiete der Physik existieren bereits Konzepttests. Der wohl bekannteste und auch am häufigsten verwendete Konzepttest ist das Force Concept Inventory (FCI) von Hestenes, Wells & Swackhamer (1992), welches das Verständnis von Konzepten der klassischen Newtonschen Mechanik erfasst. Für den FCI besteht auch eine Fassung für den deutschen Sprachraum (Gerdes & Schecker, 1999). Andere Testinstrumente für den Bereich der Mechanik und Kinematik sind die Force and Motion Conceptual Evaluation (FMCE) und der Energy and Motion Conceptual Survey (EMCS). Yeo und Zadnick (2001) entwickelten für den Bereich der Wärmelehre den Introductory Thermal Concept Inventory (ITCI), der von Engelke ins Deutsche übersetzt wurde. Des Weiteren existieren im englischsprachigen Raum Konzepttests für die folgenden Bereiche (zit. n. Lindell, Peak & Foster, 2007): Astronomie (Astronomy Diagnostic Test, ADT), Elektrizität und Magnetismus (Brief Electricity and Magnetism Assessment, BEMA; Conceptual Survey in Electricity and Magnetism, CSEM; Diagnostic Exam Electricity and Magnetism, DEEM; Determining and Interpreting Resistive Electric Circuits Test, DIRECT) sowie für Konzepte im Bereich der Wellenlehre (Wave Concept Inventory, kurz WCI).

Auch im Bereich der (Strahlen-) Optik liegen eine Reihe von Wissens- und Verstehensfragen sowie einige Tests vor, die auf die Erfassung von konzeptuellem Wissen und Verständnis in der (Strahlen-) Optik zielen und hierbei weitverbreitete Schülervorstellungen und Lernschwierigkeiten einbeziehen. Für die folgende Diskussion und den darauf aufbauenden Konzepttest wurde eine systematische Recherche mit Fokus auf das Thema Lichtausbreitung, Bildentstehung und Strahlenmodell/Bildkonstruktion durchgeführt (Bibliographie von Duit (2009), Recherche in ERIC, Google Scholar und andere Datenbanken).

Wiesner (1986; 1992a, 1992b) entwickelte mehrere Items zur gerichteten Reflexion, die auf ein adäquates Verständnis der geradlinigen Lichtausbreitung und der physikalischen Sehvorstellung zielen, nach der Licht vom Gegenstand in das Auge des Betrachters fallen muss.

Eine Zusammenstellung von Testaufgaben auf Basis der Arbeiten von Jung (1981) und Wiesner (a. a. O) findet sich in der Arbeit von Herdt (1990), allerdings ohne weitergehende psychometrische Charakterisierung.

Ebenfalls kann eine solche Zusammenstellung, ergänzt um einige weitere Aufgaben zum Thema gerichtete Reflexion, geradlinige Lichtausbreitung und Streuung von Schecker (ohne Jahresangabe) auf der Homepage des Instituts für Didaktik der Naturwissenschaften der Universität Bremen abgerufen werden (http://www.idn.uni-bremen.de/schuelervorstellungen/**)**.

Fetherstonhaugh und Treagust (1992) entwickelten einen Fragebogen über „Licht und seine Eigenschaften” (light and its properties), der aus 16 Items besteht, wobei 12 Items im Multiple-Choice-Format mit drei oder vier Distraktoren vorliegen. Die übrigen Items nutzen offene Antwortformate (vgl. ebd., S. 657). Die Autoren berichten Ergebnisse von 83 Mittelstufenschüler/-innen aus städtischen und ländlichen Gebieten Australiens, wobei die Itemschwierigkeiten und Mittelwertvergleiche in Form von t-Tests zwischen den aus der Stadt und vom Land stammenden Stichproben berichtet werden. Die Angabe psychometrischer Eigenschaften wie weiterer Itemstatistiken oder Angaben zur Prüfung der Dimensionalität des Testinstruments bleiben jedoch aus.

Sokoloff (2006) erstellte im Rahmen des Projekts „Active Learning in Optics and Photonics“ (ALOP) Multiple-Choice-Items zur Strahlen- und Wellenoptik. Das Manual, in dem die Items publiziert sind, enthält zwar eine Anleitung zum Einsatz des Tests in Schule und Unterricht (Sokloff, 2006, S. 227) jedoch keine Angaben zu empirischen Kennwerten der Items.

Bardar, Prather, Brecher und Slater (2007) entwickelten ein Testinstrument inklusive psychometrischer Angaben zu den entwickelten Aufgaben zur Erfassung von Konzepten im Bereich von Licht und Spektoskropie, den „Spectroscopy Concept Inventory (LSCI)“. Dieser Test wurde mit 548 Studienanfängern im Bereich der Astronomie erprobt und erfragt optische Konzepte, die weit über

die Inhalte der Mittelstufe hinausgehen, wie das Verständnis von Licht als elektromagnetischer Welle und seinen diesbezüglichen Eigenschaften, schwarzen Strahlern und der Dopplerverschiebung.

Chu, Treagust und Chandrasegaran (2009) entwickelten ebenfalls einen Multiple-Choice-Test, der darauf zielt Grundkenntnisse zur Lichtausbreitung bei Tag und Nacht, Lichtquellen, Wahrnehmungsprozesse und Sehen zu erheben, wobei der Test mit zweiteiligen Items arbeitet („two-tier instrument"). Dabei wird außer nach dem Vorliegen bestimmter Vorstellungen auch nach einer Begründung gefragt (mit aus der Literatur bekannten Prä- und Fehlkonzepten als Distraktoren). Die Autoren berichten einen generellen Wert der internen Konsistenz (Cronbachs α) von $\alpha_c = .65$ und geben den jeweiligen prozentuellen Anteil von insgesamt 1775 koreanischen Schüler/-innen für die Jahrgangsstufen 8 bis 10 an, welche die Items korrekt beantwortet haben (ebd., S. 258). Des Weiteren untersucht die Studie mittels multivariater Varianzanalyse und Korrelationen den Zusammenhang zwischen den Testwerten und der Einstellung zu naturwissenschaftlichen Konzepten, welche durch den Science Attitude Questionnaire (SAQ) (ebd., S. 262) erfasst wurden. Die Autoren berichten einen mittleren Zusammenhang von $r = .29$. Angaben zu den korrigierten Trennschärfen und eine Analyse der Dimensionalität erfolgen nicht.

Haagen-Schützenhöfer und Hopf (2018) schreiben ebenfalls, dass für den Bereich der Optik kein weitverbreitetes Standardinstrument publiziert wurde. Die genannten Autoren berichten auch über die Entwicklung eines Testinstrumentes mit zweiteiligen Aufgaben (Haagen-Schützenhöfer und Hopf 2012, 2014a-c). Die Gesamtreliabilität des Instrumentes wird als zufriedenstellend berichtet ($\alpha_c = 0.77$; Haagen-Schützenhöfer und Hopf, 2014a), allerdings ist das Instrument bislang nicht publiziert, sondern nur einzelne Items (z.B. Haagen-Schützenhöfer und Hopf, 2014b zu Lichtausbreitung und Sehvorgang). Auch eine Item-Analyse liegt in den veröffentlichten Arbeiten nicht vor.

Zusammenfassend kann festgehalten werden, dass etliche Autoren intensive Arbeiten für die Entwicklung von Items im Bereich der Strahlenoptik geleistet haben. Bislang fehlt jedoch ein publiziertes Testinstrument mit psychometrischer Charakterisierung, mit dem verschiedene Wissensfacetten im Bereich der Strahlenoptik in ökonomischer und psychometrisch zufriedenstellender Weise geprüft werden können. Das hier vorgestellte Testinstrument „KTSO-A" (Konzepttest Strahlenoptik – optische Abbildungen) zielt darauf, dass konzeptuelle Wissen und Verständnis von Mittelstufenschülerinnen und -schülern im Bereich der Strahlenoptik mit dem Fokus optische Abbildungen in ökonomischer Weise zu erheben und soll einen Beitrag dazu leisten, diese Lücke zu schließen. Die Testdurchführung erfordert maximal 15 Minuten, die Auswertung erfolgt auf Basis dokumentierter Lösungen; eine Lösungsschablone kann auf Anfrage zur Verfügung gestellt werden. Die Formulierung der Distraktoren basiert hierbei auf Schülervorstellungen (vgl. Liu, 2012; Lindell, Peak und Foster, 2007, S. 3), die in empirischen Studien berichtet wurden. Wir stellen deshalb im Folgenden Ergebnisse aus Forschungsarbeiten für den hier interessierenden Inhaltsbereich zusammen.

2.2 Empirische Befunde und Recherche

Fehlerhafte Schülervorstellungen im Kontext der Strahlenoptik und Mittelstufenphysik sind unter anderem für folgende Wissensbereiche dokumentiert: Vorstellungen zum physikalischen Sehvorgang, Entstehung von Licht und Schatten, Farben, Entstehung von Spiegelbildern und Entstehung reeller und virtueller Bilder bei der Sammellinse sowie allgemeine ontologischen Grundannahmen, dazu was „Licht" unter physikalischer Perspektive sein könnte (vgl. Hettmannsperger, 2015, S. 107 f.). Eine der ersten und maßgeblichen Studien zu naiven Vorstellungen über Licht stammt von Guesne (1985). Sie interviewte 30 Jugendliche im Alter von 13 und 14 Jahren, die zuvor keinen Optikunterricht in der Schule erhalten hatten, in standardisierten Interviews. Im Ergebnis zeigte sich, dass die befragten Jugendlichen Licht häufig mit seiner Quelle oder seinen Wirkungen gleichsetzten. So lokalisierten die Teilnehmenden Licht in der Quelle, z. B. in der Glühlampe. Die Gleichsetzung mit einer Wirkung fand sich beispielsweise in der Vorstellung, man könne Licht nur an Stellen sehen, auf denen das Licht auf der Wand helle Flecken erzeugt (z. B. durch Sonnenlicht oder durch die Reflexion eines Spiegels). Als zentrales Forschungsergebnis hob die Autorin hervor, dass kaum einer der Befragten erkannte, dass Gegenstände, die nicht selbst leuchten, unter der Beleuchtung durch Tageslicht oder Lampen Licht zurückwerfen. Daraus resultieren Defizite beim Verständnis des Sehvorgangs. Da die befragten Kinder im Alltag Licht nur erkennen, wenn es einen deutlich wahrnehmbaren Effekt hervorbringt, glaubten sie nicht, dass auch bei nicht-selbst-leuchtenden Körpern Licht ins Auge gelangt (vgl. Guesne, 1985, S. 91).

Auf Basis dieser und ähnlicher Beobachtungen stellte Wiesner in einem mehrjährigen Forschungs- und Entwicklungsprogramm weitere zentrale Lernschwierigkeiten in der Strahlenoptik zusammen (Wiesner, 1988; 1992a; 1992b; 1994). Dabei wird insbesondere der enge Zusammenhang zwischen einer fehlerhaften physikalischen Sehvorstellung und fehlerhaften Konzepten zur Streuung deutlich (Wiesner, 1992a): So gehen viele Schülerinnen und Schüler (wie in der Studie von Guesne, s. o.) davon aus, dass beleuchtete Gegenstände wie Tische, Bücher oder Bilder kein Licht abstrahlen (Wiesner, 1986). Auch hier fehlt ein klares Gesamtbild von der Lichtausbreitung von der Quelle über die Streuung an beleuchteten (und daher sichtbaren) Gegenständen bis zum Auge des Beobachters (von Wiesner „Sender-Empfängervorstellung" genannt). In den genannten Arbeiten wird insbesondere herausgearbeitet, dass ein solches Gesamtbild notwendige Voraussetzung für das Verständnis von Strahlengangkonstruktionen ist, wie sie z. B. bei der in dieser Arbeit interessierenden optischen Abbildung zugrunde liegen. Die genannte Gruppe von Lernschwierigkeiten schließt das Fehlkonzept ein, das auftreffende Licht mache die Gegenstände hell, bleibe auf diesen liegen oder verschwinde allmählich (ebd.). Ebenfalls damit verbunden ist die Vorstellung, dass von Lichtquellen mit geringer Intensität, wie Räucherstäbchen oder weit entfernte beleuchtete Fenster, kein Licht ins Auge gelangt (ebd.). Spätere Studien bestätigen, dass die physikalische Erklärung des Sehvorgangs vielen Schüler/-innen Schwierigkeiten bereitet (Selley, 1996; Chu et al., 2009).

Die Schwierigkeiten mit der physikalischen Sehvorstellung schlagen sich nach Wiesner (1992a) auch auf Lernschwierigkeiten mit dem Spiegelbild nieder. Auch hier erkennen viele Schülerinnen und Schüler nicht, dass das Licht aus der Richtung des Spiegels ins Auge fallen muss, damit das Spiegelbild wahrgenommen werden kann. Schülerinnen und Schülern bereitet 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 Schülerinnen und Schülern als nicht nachvollziehbar eingestuft. Die Schülerinnen und Schüler gehen oft davon aus, dass das Spiegelbild auf der Spiegeloberfläche liegt (vgl. Wiesner, 1986, S. 26, S. 27.; 1992a; 1992b, S. 288). Der Spiegel wird auch oft als ein Gegenstand aufgefasst, der das Spiegelbild zum Betrachter zurückwirft. Auch mit der Entstehung von reellen Bildern durch die Sammellinse ist gemäß Wiesner (1994, S. 8) eine Reihe von Fehlvorstellungen verbunden: Viele Schülerinnen und 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 geht als Ganzes durch die Linse zum Schirm und wird dabei in der Linse umgedreht (Wiesner, 1994, S. 8). Wiesner 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, dass das Bild als Ganzes vom Gegenstand aus durch die Linse auf den Schirm transportiert wird, ist es nur konsequent anzunehmen, dass ein Teil des Bildes abgeschnitten wird, 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, dass das Bild kreisförmig am äußeren Rand abgeschnitten wird. Wird die Linse zur Hälfte abgedeckt, gehen viele Schülerinnen und Schüler aber auch Studierende davon aus, dass auch das reelle Bild zur Hälfte abgeschnitten wird, einige überlegen sich sogar, welche Hälfte des Bildes (obere versus untere Hälfte) betroffen ist (vgl. Goldberg & McDermott, 1987, S. 112; Wiesner, 1994, S. 8).

Weitere Vorstellungen, die ebenfalls das Konzept der Punkt-zu-Punkt-Abbildung außer Acht lassen, bestehen in der Idee, die Linse konzentriere das Licht, oder hinter der Linse seien mehr Licht bzw. mehr Strahlen vorhanden als vor der Linse (vgl. Wiesner, 1986). Häufig wird die Entstehung reeller Bilder durch Spiegelung und Reflexion erklärt, dabei wird einem Gegenstandspunkt in der Regel nur ein Strahl zugeordnet und nicht ein divergierendes Strahlenbündel (vgl. ebd.).

Goldberg und McDermott (1987) berichten über das Verständnis der Entstehung reeller Bilder durch Sammellinse und Hohlspiegel von der folgenden weiteren Verständnisschwierigkeit bei Collegestudent/-innen (80 Studierende aus einem Einführungskurs zur Physik). Diesen wurde ein Versuchsaufbau gezeigt, bei dem eine Glühlampe, eine Linse und ein Schirm hintereinander auf einer optischen Bank montiert sind. Im Verlauf des Interviews wurden die Studierenden gefragt, wo das Bild wäre, wenn man den Schirm entfernt und sie frei um den Versuchsaufbau im Raum herumgehen können (natürlich wird hier von den Autoren der physikalische Bildbegriff als Schnittpunkt von Lichtstrahlen und als Ort des Luftbildes bei der reellen Abbildung vorausgesetzt, siehe z. B. Meschede (2006, S. 484)). Nur wenige Studierende waren in der Lage zu erkennen, dass sich das Bild an der gleichen Position befindet wie zuvor der Schirm. Die übrigen Studierenden gaben eine Erklärung ab wie etwa, dass sich das Bild auf oder in der Linse befindet. Insbesondere war die Vorstellung verbreitet, dass ein Bild nur mit Hilfe eines Schirms gesehen werden kann und dass die Linse das Bild quasi einrahmt (vgl. Goldberg & McDermott, 1987, S. 114). Auch Guesne (1985) hatte bereits in ihrer oben genannten Studie Kinder zur Rolle der Sammellinse in der Funktion als Lupe und Brennglas befragt. Dabei konnte sie zwei Antworttypen auffinden: Die erste Hälfte der Antworttypen beinhaltet die Ansicht, das Vergrößerungsglas „vermehrt“ das Licht, während die andere Hälfte der Antworten die Überzeugung umfasst, dass die Sammellinse das Licht konzentriert. Kinder, welche das Konzept der Lichtkonzentration vertraten, waren der Ansicht, die gesamte Lichtmenge, die durch das Vergrößerungsglas hindurchgehe, bleibt hinter der Linse erhalten, was wissenschaftlich korrekt ist. Dass auch diese Gruppe von Kindern nicht notwendigerweise eine physikalisch angemessene Vorstellung von der Funktionsweise hat, zeigte sich in den Zeichnungen der Kinder, mit denen sie verdeutlichen sollten, wie die Linse das Licht konzentriert (vgl. Guesne, 1985, S. 87). Kinder, die der Ansicht waren, das Vergrößerungsglas „vermehrt“ das Licht, gaben entweder an, dass hinter der Lupe mehr Licht ist als vor der Lupe oder dass das Licht hinter der Lupe verstärkt bzw. vermehrt wird.

Die Literatur zeigt demnach, dass Lernende Vorstellungen von Licht besitzen, welche sich grundlegend von einem wissenschaftlichen Verständnis unterscheiden. Dies offenbart sich insbesondere auch in substanzbasierten Vorstellungen von Licht. Reiner et al. (2000, S. 14, 15) beschreiben eine solche Konzeption wie folgt: Wenn Lernende gefragt werden, wie Sehen funktioniere, gaben sie an, dass Moleküle, im Sinn von Partikeln, zwischen dem gesehenen Gegenstand z. B. einem „Buch“ und dem „Auge“ vorhanden seien. Der Sehvorgang wird demgemäß als das Ergebnis beweglicher Lichtpartikel interpretiert. Das scheint zwar im Einklang mit dem Konzept von Photonen zu sein, Licht wird dabei aber oft als Flüssigkeitsstrom beschrieben, der in Bewegung ist, sich aber auch in Ruhelage befinden kann. Damit gehen Schwierigkeiten einher, Lichtstrahlen als Modellvorstellung zu erkennen und zwischen den Begriffen Lichtbündel und Lichtstrahl differenzieren zu können. Der Vollständigkeit halber sei hier auch auf neuere Arbeiten zum konzeptuellen Verständnis von Farben hingewiesen (Martinez-Borreguero et al., 2013; Naranjo-Correa et al., 2015); dieses Gebiet liegt jedoch außerhalb der Domäne des hier vorgestellten Tests.

Tab. 1. Angaben zu Items des KTSO-A

Konzept	Distraktoren: Schülervorstellungen mit Beispielen	Item/s	Quellen
Sekundäre Lichtquellen und physikalische Sehvorstellung, Streuung	Unterscheidung primäre und sekundäre Lichtquellen In einem vollständig abgedunkelten Raum sind helle Gegenstände sichtbar. Bei Lichtquellen mit geringer Intensität gelangt kein Licht mehr ins Auge.	1	Wiesner, 1992a, S. 16
Lichtausbreitung	Streuung und (diffuse) Reflexion an Oberflächen Beleuchtete Gegenstände wie Tische, Bücher oder Bilder strahlen kein Licht ab / beleuchtete Gegenstände werfen kein Licht zurück. Licht, das auf eine Oberfläche fällt, bleibt auf dieser liegen und macht sie hell.	2, 4, 5	Wiesner, 1986, S. 26, 27
Ontologisches Verständnis von Licht (-strahlen) und Modell-Kompetenz	Licht und Lichtstrahlen als ontologische Kategorie Licht wird als Substanz aufgefasst: Licht wird als Flüssigkeitsstrom beschrieben, der in Bewegung ist, sich aber auch in Ruhelage befinden kann. Verwechslung von Lichtbündel und Lichtstrahl	3	Reiner et al., 2000, S. 15ff.
Entstehung reeller Bilder bei der Sammellinse	Linsenabbildung / Bildentstehung 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. Die Linse konzentriert oder „vermehrt" Licht	6,7	Abbildungen zu Item 6 und 7 in Anlehnung an Wiesner, 1986, S. 28
Entstehung reeller Bilder bei der Sammellinse	Fehlvorstellung „Blende schneidet Bild ab" Bei Abdeckung mit einer ringförmigen Blende werden die Bildränder kreisförmig abgeschnitten. Bei Abdeckung der oberen bzw. unteren Hälfte der Linse wird entsprechend die obere bzw. untere Hälfte des Bildes angeschnitten.	8a, 8b, 8c	weitere Beschreibung der Schülervorstellungen Wiesner, 1994, S. 8

2.3 Rahmen und inhaltliche Spezifikation

Auf der Basis der zuvor dargestellten Arbeiten zu Schülervorstellungen im Bereich der Strahlenoptik zielt das hier vorgestellte Testinstrument darauf, das konzeptuelle Verständnis im Bereich der Strahlenoptik mit dem Schwerpunkt optische Abbildungen bei der Sammellinse und hiermit verbundener physikalischer Grundkonzepte zu erfassen. Der Test wurde im Rahmen des Projekts „Das Experiment als Mittel zur Entwicklung von Repräsentationskompetenz im Rahmen einer problemorientierten Aufgabenkultur des Physikunterrichts" (2009 bis 2012) des DFG-Graduiertenkollegs „Unterrichtsprozesse" (GK 1561) der Universität Koblenz-Landau am Campus Landau entwickelt, um mögliche Lernfortschritte in Bezug auf Grundkonzepte in der Strahlenoptik in zwei aufeinander bezogenen Interventionsstudien Scheid (2013) und Hettmannsperger (2015) analysieren zu können. Die verwendeten Items erfassen Grundkonzepte aus den Bereichen Lichtausbreitung, Streuung, Reflexion und Linsenabbildung (vgl. Tab. 1). Andere Bereiche, wie z. B. Licht und Farben oder Licht und Schatten sowie Spiegelbilder, welche für das Verständnis optischer Abbildungen keine oder eine vergleichsweise untergeordnete Rolle spielen, wurden nicht in den Test aufgenommen. Wie die Literatur und die nachfolgende vorgestellte empirische Analyse (siehe Kapitel 3 und 4) zeigen, umfasst der Phänomenbereich *optische Abbildung bei der Sammellinse* (also das Zustandekommen eines reellen Bildes) einen hinreichend großen und inhaltlich zusammenhängenden Teil der Strahlenoptik, um die Entwicklung eines Konzept-Tests für diesen Bereich zu rechtfertigen (siehe auch Hettmannsperger, 2015, S. 129-131).

Das Vorgehen der Itemkonstruktion entsprach im ersten Schritt der rationalen Testkonstruktion (vgl. Bühner, 2011, S. 93) und der Literatur zur Standards bei der Entwicklung von Konzepttests und anderen diagnostischen Tests (Adams & Wiemann, 2011, Madsen et al, 2017): Auf Basis vorhandener Literatur zu Schülervorstellungen und Lernschwierigkeiten und in der Literatur beschriebener Aufgaben wurden zu jedem jeweilig gewünschten wissenschaftlichen Konzept, eine korrekte Antwortalternative und Distraktoren entwickelt (vgl. Tabelle 1). Die Konstruktion der Distraktoren folgt daher forschungsgeleitet (vgl. Haladyna & Downing. 1989, S. 68 zur methodischen Begründung des Vorgehens und Liu, 2012; Lindell, Peak und Foster, 2007, S. 3 zum Vergleich mit Arbeiten anderer Autoren). Aus diesem Arbeitsschritt resultierten zunächst 21 Items, die einer empirischen Überprüfung (Itemanalyse und Exploratorische Faktorenanalyse) unterzogen und eliminiert wurden, wenn die Kennwerte außerhalb der etablierten Akzeptanzbereiche lagen (Ding & Beichner, 2009). Detaillierte Angaben zu diesem Überarbeitungsprozess finden sich in Hettmannsperger (2015, S. 132 f.). Die resultierende Testfassung enthält 10 Items mit zufriedenstellenden bis guten psychometrischen Eigenschaften, die im Folgenden berichtet werden.

3 STICHPROBE UND METHODEN

3.1 Stichprobe

Der Test wurde mit n = 869 Schüler/-innen der Klassenstufen 7 und 8 aus 38 Klassen an 15 Gymnasien und 3 Gesamtschulen erprobt (vgl. Tabelle 2); die Schüler/-innen waren im Schnitt 13 Jahre alt (M = 13.4 Jahre, SD = 0.7). Alle Schüler/-innen erhielten erstmals Unterricht zum Thema „Strahlenoptik". Zur Itemanalyse wurden Daten aus den zwei oben genannten Studien genutzt (Hettmannsperger, 2015; Scheid, 2013). Dabei war mit Blick auf die bekanntermaßen hohen kognitiven Anforderungen

der Nutzung von multiplen Repräsentationen eine Entscheidung für eine mehrheitlich aus dem Gymnasium stammenden Stichprobe getroffen worden, um zunächst bei Schülern mit günstigen Lernvoraussetzungen die Wirksamkeit der untersuchten Interventionen zu erheben. Weitergehende Angaben zu der Art der Interventionen (die hier nicht Gegenstand sind) und zur Stichprobe finden sich in Hettmannsperger (2015, S. 143 f.) und Scheid (2013, S. 93 f.; Scheid et al., 2019).

Die Teilstichproben werden in Abhängigkeit davon, ob eine gezielte Intervention für die Förderung des konzeptuellen Verständnisses und eine Überwindung von Lernschwierigkeiten und Fehlvorstellungen im Themenbereich des Tests stattfand, wie folgt unterschieden: Nur in Teilstichprobe A fand die vorgenannte Intervention statt, in Teilstichprobe B fand eine andere Intervention statt (vgl. Scheid et al., 2019). Gemäß Bühner (2011, S. 81) kann die Größe der Stichproben als gut bis sehr gut angesehen werden.

Tab. 2. Stichprobenangaben

Level	*N* Gesamt	*n* Studie A	*n* Studie B
Schulen	18	10	8
Klassen	*38*	*21*	*17*
Schüler/-innen	869	480	389

Studie A – Hettmannsperger, 2015; Studie B - Scheid, 2013

Stichprobe A (Hettmannsperger, 2015): Der Unterricht zielte in beiden Bedingungen (Treatment- und Kontrollgruppe) auf die Förderung des konzeptuellen Verständnisses und auf eine Überwindung von Lernschwierigkeiten und Fehlvorstellungen im Themenbereich dieses Artikels. Treatment- und Kontrollgruppe (TG und KG) dieser Studie unterschieden sich dahingehend, ob die Schülerinnen und Schüler kognitiv aktiviert wurden sich mit multiplen Repräsentationen auseinander zu setzen (TG) oder keine solche auf multiple Repräsentationen gezielte kognitive Aktivierung stattfand (KG), während sie in beiden Gruppen Aufgaben bearbeiteten, die Schülervorstellungen thematisierten. Da zwischen den beiden Bedingungen keine signifikanten Unterschiede bestanden (Hettmannsperger, 2015, S. 146 f.), werden im Folgenden die Werte der Stichprobe A insgesamt berichtet und nicht getrennt nach Bedingungen.

Teilstichprobe A kann also exemplarisch zeigen, ob der Test in der Lage ist, die Wirksamkeit einer auf die entsprechenden Lernschwierigkeiten/Fehlkonzepte zielenden Intervention nachzuweisen. Außerdem dient diese Stichprobe dazu, eine etwaige Substruktur des Tests per Kreuzvalidierung zu prüfen (s. Abschnitt 3.4).

Stichprobe B (Scheid, 2013): Der Unterricht in dieser Stichprobe enthielt keine gezielte Intervention zur Überwindung von Lernschwierigkeiten und Fehlvorstellungen im Themenbereich dieses Artikels. Gegenstand des Vergleichs zwischen TG und KG war vielmehr die Förderung der Fähigkeit, Kohärenz zwischen mehreren Repräsentationen herzustellen (siehe auch Scheid et al., 2019). Da zwischen den beiden Bedingungen bezüglich des Konzeptverständnisses keine signifikanten Unterschiede bestanden (Scheid, 2013) werden im Folgenden die Werte der Stichprobe B ebenfalls gesamt berichtet und nicht getrennt nach Bedingungen.

Teilstichprobe B entspricht also der üblichen Situation für eine Vergleichsstichprobe für Konzept-Tests: nach Unterricht, aber ohne gezielte Maßnahmen hinsichtlich der fraglichen Lernschwierigkeiten und Schülervorstellungen.

3.2 Durchführung und empirisches Vorgehen

Die vorgestellten Ergebnisse basieren auf den Daten von Schülerinnen und Schülern, die erstmals in Klasse 7 oder zu Beginn von Klasse 8 Optik Unterricht zum Thema Strahlenoptik erhalten hatten (s. o.). Die Darstellung der Ergebnisse dieses Messzeitpunkts wurde gewählt, da sie dem Kenntnisstand der Schüler/-innen der Mittelstufe nach Unterricht (laut Lehrplan des entsprechenden Schultyps) entspricht. Die vollständige Testfassung ist in Appendix 2 dieses Artikels abgedruckt. Die Schülerinnen und Schüler benötigten in der Regel nicht mehr als 10 – 15 Minuten, um den Test zu bearbeiten. Die Schüler/-innen wurden in einer standardisierten kurzen Testinstruktion darauf hingewiesen, dass etwaige Abbildungen Bestandteil der Aufgaben seien, und für eine Lösung sorgfältig zu beachten seien, und dass manchmal entweder nur eine Antwort oder manchmal mehrere Antworten richtig sein können.

Die Bewertung der Aufgaben wurde wie folgt vorgenommen: Die Antworten wurden mit 2 Punkten bewertet, wenn ausschließlich die korrekte Alternative gewählt wurde, ein Punkt wurde vergeben, wenn zusätzlich zu der richtigen Antwort auch ein Distraktor angekreuzt wurde (2 Punkte für die Lösung weniger 1 Punkt Abzug). In allen anderen Fällen wurde die jeweilige Antwort mit 0 Punkten bewertet.

Im Sinn der klassischen Testtheorie (vgl. Lienert & Raatz, 1998; Rost, 2004; Bühner, 2011) wurden Itemkennwerte wie Lösungswahrscheinlichkeit und korrigierte Trennschärfe sowie interne Konsistenz (α_c) als Schätzer für die Reliabilität berechnet. Des Weiteren wurde eine deskriptive Analyse der Distraktoren durchgeführt. In Anbetracht der Anzahl der Distraktoren (k = 3) (Wilcox. 1981) und dem Messzeitpunkt (das Verständnis der Schülerinnen und Schüler wurde nach dem Unterricht erfasst), kann für einen gegebenen Distraktor eine Untergrenze von mindestens 5 % an Antworten als zufriedenstellend betrachtet werden (vgl. auch Wakefield. 1958 zit. n. Haladyna & Downing, 1989, S. 58; Kline, 2005, S. 57).

Unter Berücksichtigung des Skalenniveaus wurde für die Berechnung der Trennschärfen und α_c polychorische bzw. polyseriale Korrelationen verwendet (vgl. Eid, Gollwitzer & Schmitt, 2011, S. 538). Die Berechnung erfolgte mit Hilfe des R-Pakets „polycor“ (Fox, 2010). Polyseriale Korrelationen sind für die Berechnung der Korrelation zwischen ordinalskalierten und intervallskalierten Daten die Methode der Wahl (vgl. Eid et al., 2011, S. 538), polychorische Korrelationen für die Berechnung von Korrelationen zwischen ordinalskalierten Daten untereinander (vgl. ebd., 2011, S. 515).

Zur Schätzung, wie gut sich die jeweiligen Items dazu eignen, zwischen Personen mit unterschiedlichen Fähigkeiten zu differenzieren, wurden die korrigierten Eigentrennschärfen, d. h. die Korrelation des jeweiligen Items

mit der Gesamtpunktzahl des Tests (ausgenommen des jeweiligen Items selbst) berechnet. Da Varianzen bei Daten mit drei Abstufungen nur begrenzt interpretierbar sind, wurde das standardisierte Cronbach's Alpha verwendet, welches auf den durchschnittlichen Korrelationen zwischen den Items basiert (Bonanomi, Nai Ruscone & Osmetti, 2013).

Aus zwei Gründen wurde darauf verzichtet, den Test mittels eines Rasch-Modells zu skalieren: Zum ersten ist der Test so konzipiert, dass die Aufgaben der Skala Abdeckaufgaben sich auf den gleichen Experimentalaufbau beziehen, bei dem die Linse durch eine Blende (teilweise) abgedeckt wird. Aus diesem Grund ist die Voraussetzung der lokalen stochastischen Unabhängigkeit der Aufgaben nicht erfüllt (vgl. Strobl, 2012, S. 17). Gegen ein klassisches eindimensionales Raschmodell spricht des Weiteren zudem, dass der Test nicht eindimensional ist, da mindestens die Aufgaben zur Abdeckung der Sammellinse eine Subdimension des Tests bilden sollten.

Es wurde zunächst auch geprüft, ob der Test tatsächlich wie erwartet mehrdimensional ist, anschließend wurde das Instrument mittels exploratorischer und konfirmatorischer Faktorenanalyse kreuzvalidiert (Bühner, 2011; Anwendungen auf Konzepttests: Ramlo, 2008; Scott, Schuhmayer & Gray, 2012).

Wie bekannt, ermöglicht eine exploratorische Faktorenanalyse eine Datenstruktur zu finden (Bühner, 2011, S. 254). Deshalb wurde zunächst eine exploratorische Faktorenanalyse (EFA) auf Basis der Stichprobe B durchgeführt. Das Modell, welches sich aus der exploratorischen Faktorenanalyse ergab, wurde dann mittels der Methode der konfirmatorischen Faktorenanalyse anhand der Stichprobe A überprüft (ebd.). Abschließend wurden diese Schritte vice versa wiederholt, um sicherzustellen, dass sich für beide Teilstichproben die gleiche Teststruktur findet. Dies entspricht dem Verfahren der Kreuzvalidierung (ebd.).

Die exploratorische Faktorenanalyse wurde mit der freien Statistiksoftware R unter Verwendung der Pakete psych (Revelle, 2013), lavaan (Rosseel, 2012) und qgraph (Epskamp et al., 2012) durchgeführt. Der Vorteil der Statistiksoftware R besteht darin, dass sie die Möglichkeit bietet, der Analyse eine polychorische Korrelationsmatrix für ordinalskalierte Daten zugrunde zu legen. Eine Normalverteilung der Daten ist hierbei keine zwingende Voraussetzung (Bühner, 2011).

Abschließend wurde ein Expertenrating mit 11 Experten aus dem Bereich Physikdidaktik zur Beurteilung der internen Validität des Tests durchgeführt und die Korrelationen mit den Schulnoten in Mathematik und Physik angegeben. Weitere Angaben zur Expertenstichprobe und den Ratingfragen finden sich in Abschnitt 4.3.

4 ERGEBNISSE

4.1 Item- und Teststatistiken

Die Analyse der Itemschwierigkeiten (vgl. Abb. 1 und Tab. 3) ergab, dass für die Gesamtstichprobe und die Stichprobe A alle Items innerhalb des Toleranzbereichs von $0.20 \leq P_i \leq 0.80$ liegen (Kline, 2015). In der Stichprobe B bestehen kleinere Abweichungen für Item 2, das Item ist etwas zu „leicht" sowie für die Items 8a-c, diese Items sind etwas zu schwer. Für die Trennschärfen ergeben sich in der Gesamtstichprobe Werte in einem Bereich zwischen $.23 \leq r_{it} \leq .43$ (vgl. Abb. 2 und Tab. 3). Auch diese Werte liegen innerhalb des Toleranzbereichs von $r_{it} \geq 0.2$ (Kline, 2015). Für A sind die Werte durchwegs höher, für B liegen die Trennschärfen bei etlichen Items nur bei knapp 0.20 (s. Diskussion).

Auch die Werte für interne Konsistenz (α_C; Tab. 3) als Schätzer für die Reliabilität des Tests können gemäß gängiger Richtlinien (COTAN-System, Evers, 2001; darauf beruhend: EFPA, 2013) als ausreichend für Gruppenvergleiche angesehen werden), der Wert für die Teilstichprobe A ist nahe dem Schwellenwert für Individualdiagnostik (Bühner, 2011; Ding & Beichner, 2009; Doran, 1980). Die Werte in Tabelle 3 belegen zudem, dass die interne Konsistenz weder für die Teilstichproben noch für die Gesamtstichprobe durch den Ausschluss weiterer Items angehoben werden könnte.

Die Distraktorenanalyse, welche in Tabelle 7 berichtet wird, zeigt auf, dass für jedes Item mehr als 5 % der Antworten auf mindestens einen der Distraktoren fällt. Unter Berücksichtigung der o. g. Schwelle für 5 % bei 3 Distraktoren (k = 3) (Wilcox. 1981; Haladyna & Downing, 1989, S. 58; Kline, 2005, S. 57) und dem Messzeitpunkt (das Verständnis der Schülerinnen und Schüler wurde nach dem Unterricht erfasst), muss keines der Items aufgrund eines zu selten gewählten Distraktors ausgeschlossen werden.

4.2 Analyse der Dimensionalität

Zur Analyse der Dimensionalität wurde zunächst faktorenanalytisch geprüft, ob alle Items auf einen Generalfaktor laden. Erwartungskonform zeigte sich, dass das Modell mit einem Generalfaktor verworfen werden musste (χ^2 (119) = 1390, $p < 0.001$; *CFI* = 0.60; *TLI* = 0.54; *RMSEA* = 0.11; *SRMR* = 0.10). Übereinstimmend mit den theoretischen Vorüberlegungen ist daher von einem mehrdimensionalen Konstrukt auszugehen. So war anzunehmen, dass mindestens die Aufgaben zur Abdeckung der Sammellinse eine eigene Skala bilden.

Um die Struktur des Tests genauer zu analysieren, wurde der Test nun einer Kreuzvalidierung unterzogen. Hierbei sollten sich für beide Stichproben A und B die gleichen Facetten bzw. Subdimensionen ergeben. Das Vorgehen der Kreuzvalidierung umfasste drei Schritte:

1) Es wurde eine exploratorische Faktorenanalyse mit Stichprobe B umgesetzt.
2) Das ermittelte exploratorische Modellwurde mittels einer konfirmatorischen Faktorenanalyse an Stichprobe A überprüft.
3) Die Schritte 1) und 2) wurden vice versa wiederholt.

Tab. 3. Itemkennwerte

Item		Schwierigkeit			Korrigierte Trennschärfe			Interne Konsistenz α_C unter Ausschluss des jeweiligen Items		
		A	B	Gesamt	A	B	Gesamt	A	B	Gesamt
Nr.	Inhalt	n = 480	n = 389	N = 869	n = 480	n = 389	N = 869	n = 480	n = 389	N = 869
y1	*Sichtbare Gegenstände in dunklem Raum*	0.74	0.81	0.77	0.28	0.24	0.23	.78	.62	.74
y2	*Helligkeit eines Zimmers - Farbe der Tapete*	0.70	0.63	0.67	0.28	0.18	0.25	.78	.63	.74
y3	*Lichtstrahlen sind …*	0.42	0.39	0.41	0.30	0.20	0.27	.78	.63	.74
y4	*Lichtstrahl in dunklem Raum*	0.35	0.34	0.35	0.29	0.15	0.23	.78	.63	.74
y5	*Lichtstrahl: Kreidestaub*	0.68	0.76	0.71	0.41	0.30	0.33	.76	.58	.72
y6	*Entstehung reeller Bilder an Sammellinse*	0.43	0.35	0.39	0.26	0.19	0.25	.78	.62	.74
y7	*Bildkonstruktion: Bildentstehung*	0.40	0.29	0.35	0.34	0.22	0.32	.77	.62	.73
y8a	*Abdeckung einer Linsenhälfte*	0.55	0.18	0.38	0.47	0.17	0.40	.76	.62	.72
y8b	*Ringförmige Abdeckung der Linse*	0.40	0.18	0.30	0.48	0.23	0.43	.76	.60	.71
y8c	*Enge ringförmige Abdeckung der Linse*	0.39	0.17	0.29	0.48	0.19	0.42	.75	.61	.71

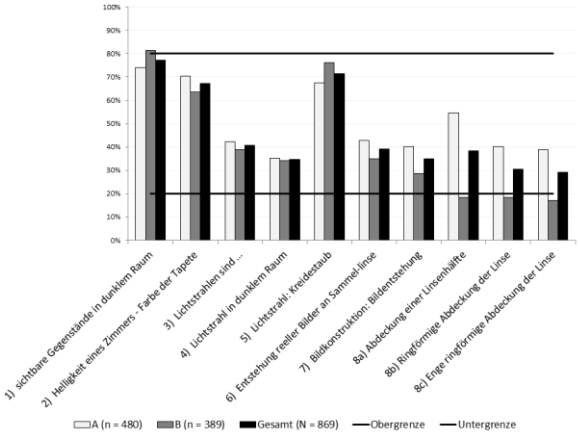


Abb. 1. Itemschwierigkeiten des KTSO-A je Stichprobe

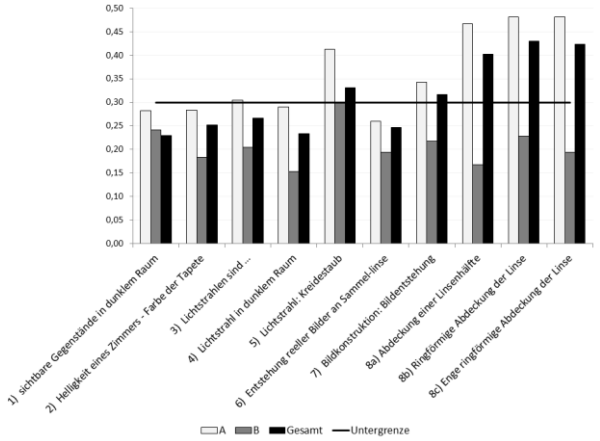


Abb. 2. Korrigierte Trennschärfen des KTSO-A je Stichprobe

Diese Abfolge sichert die Vorgehensweise insofern ab, als eine etwaige Teststruktur exploratorisch in der Teilstichprobe ohne gezielte Intervention (Stichprobe B, weniger ausgeprägte Wissensstruktur) gesucht wird und erst anschließend in Stichprobe A überprüft wird. Bei der Analyse wurde die Annahme zugrunde gelegt, dass die Faktoren untereinander korrelieren können, d.h. dass die Beantwortung von bestimmten Itemgruppen nicht notwendigerweise unabhängig von anderen Itemgruppen erfolgt. Daher wurde in der exploratorischen Faktorenanalyse gemäß der Empfehlung von Bühner (2011, S. 349) das oblique Rotationsverfahren „Promax“ vorwendet. Die Prüfung der Voraussetzungen ergab, dass die Bewertungskriterien zur Durchführbarkeit der Analyse vollständig erfüllt sind (vgl. Bühner, 2011, S. 346). So belegt der Bartlett-Test, dass die polychorische Korrelationsmatrix signifikant von der Einheitsmatrix verschieden ist ($\chi^2 = 450$, $df = 45$, $p < 0.001$) und somit faktorisiert werden kann. Der Kaiser-Meyer-Olkin-Koeffizient weist darauf hin, dass sich die Itemauswahl für die Faktorenanalyse eignet und der empirisch vorgefundene Wert von 0.79 im Sinne der üblichen Qualitätsstandards in einem guten Bereich liegt (vgl. ebd., S. 347). Auf Itemebene zeigen die MSA-Koeffzienten (Measure of Sample Adequacy) ebenfalls gute Werte $0.73 \leq MSA \leq 0.89$. Der Verlauf der Eigenwerte und das Ergebnis der Parallelanalyse nach Horn (1965) weisen darauf hin, dass die ersten drei Faktoren die Hauptinformationen der Daten repräsentieren (vgl. Abb. 3 und Tab. 5). Die Kommunalitäten liegen mit $0.11 \leq \widehat{H} \leq 0.86$ in einem für die Stichprobe von n = 389 und drei bis vier Items pro Faktor ebenfalls in einem akzeptablen bis guten Bereich (vgl. Bühner, 2011, S. 345). Es zeigt sich, dass die Items 8 a-c einen mittleren, die Items 1, 4, 5, 7 einen mittleren bis hohen und die Items 2, 3 und 6 den geringsten Varianzanteil aufklären (siehe Tab. 4 und Tab. 7). Von den Faktorladungen nach Schritt 1) sind vier Werte gut oder besser, drei angemessen, zwei akzeptabel, und zwei zu klein (Tabachnick & Fiedell, 2007).

Es ergeben sich demnach drei Dimensionen für den untersuchten Test. In Abbildung 4 wird eine Darstellung der durch die Faktorenanalyse gefundenen Ergebnisse gege-

ben, die an die Darstellungen von Strukturgleichungsmodellen angelehnt ist, um den Bezug zur konfirmatorischen Faktorenanalyse in Schritt 2) darstellen zu können. Alle drei genannten Dimensionen tragen zum Verständnis der Bildentstehung mittels Sammellinse bei, was sich in den wechselseitigen mittleren Korrelationen ($.26 \leq r \leq .42$) zeigt (siehe Abb. 4).

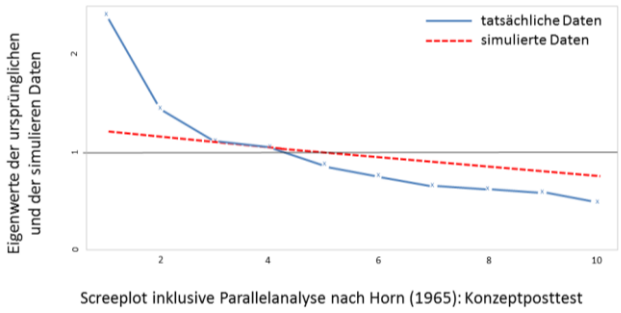


Abb. 3. Screeplot inklusive Parallelanalyse nach Horn (1965)

Insgesamt werden 33 % der Gesamtvarianz durch die drei Faktoren aufgeklärt. Im zweiten Schritt wurden die Ergebnisse der exploratorischen Faktorenanalyse mit der Stichprobe A (n = 480) kreuzvalidiert.

Da es sich um ordinale Daten handelt, wurde zur Durchführung der konfirmatorischen Faktorenanalyse die ADF-Methode verwendet, welche keine multivariat normalverteilten Daten auf Itemebene voraussetzt (vgl. Bühner, 2011, S. 432). Des Weiteren wurde geprüft, dass keine Ausreißer vorliegen und keine Kollinearität der Items ($r_{(n = 389)} < 0.76$) besteht (vgl. ebd.).

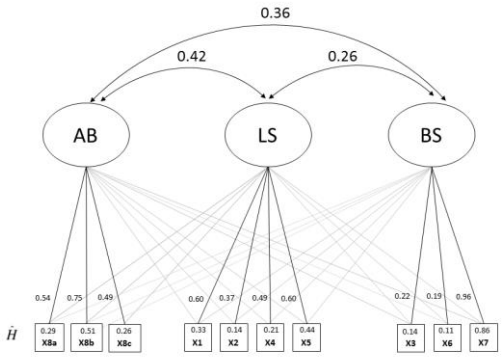


Abb. 4. Exploratorische Faktorenanalyse. AB = Bildentstehung inklusive Abdeckaufgaben, LS = geradlinige Lichtausbreitung und Streuung, BS: Bildkonstruktion / Strahlenmodell

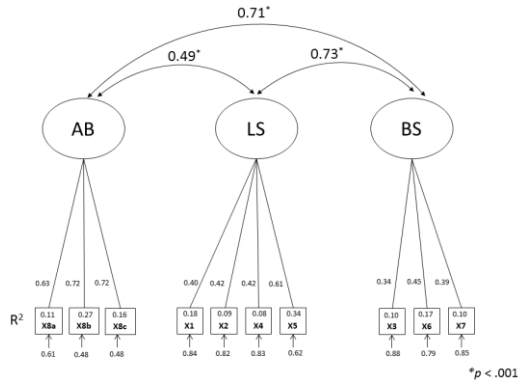


Abb. 5. Konfirmatorische Faktorenanalyse: AB = Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben, LS = geradlinige Lichtausbreitung und Streuung, BS: Verständnis Bildkonstruktion / Strahlenmodell.

Auf Basis des Modells, das sich an der exploratorischen Faktorenanalyse orientierte, werden die Items wie folgt zugeordnet (siehe Abb. 5): AB: Abdeckaufgaben mit Verständnisfragen zur Bildentstehung: y8a, y8b, y8c; LS: geradlinige Lichtausbreitung und Streuung: y1, y2, y4, y5 und BS: Verständnis zur Bildkonstruktion, zu Lichtquellen und zum Strahlenmodell: y3, y6, y7. Die Ergebnisse der konfirmatorischen Faktorenanalyse auf Basis von Stichprobe A bestätigen die vorgefundene Struktur, welche die exploratorische Faktorenanalyse aufzeigte.

Insgesamt ergaben sich gute globale Fit-Indizes ($\chi^2_{(32)}$ 44, p = 0.075; *CFI* = 0.98; *TLI* = 0.98, *RMSEA* = 0.05; *SRMR* = 0.03). Wie bei den Ergebnissen der exploratorischen Faktorenanalyse weisen die Korrelationen der Faktoren untereinander darauf hin, dass alle drei Dimensionen zum Verständnis der Bildentstehung mittels Sammellinse beitragen und wechselseitig inhaltlich miteinander zusammenhängen ($0.49 \leq r \leq 0.73$). In Bezug auf die lokalen Gütekriterien zeigte sich, dass jeder der Pfadkoeffizienten signifikant von Null verschieden war.

In Schritt drei der Analyse wurden beide Analysen exploratorische und konfirmatorische Faktorenanalyse mit der jeweils anderen Teilstichprobe durchgeführt. In beiden Fällen konnten die Ergebnisse jeweils mit geringfügigen Abweichungen repliziert werden, daher werden die Ergebnisse hier lediglich im Überblick dargestellt.

Insgesamt werden in Stichprobe A in der exploratorischen Faktorenanalyse 44 % der Gesamtvarianz durch drei Faktoren mit Eigenwerten > 1 aufgeklärt und in Stichprobe B 33 % der Gesamtvarianz. In Stichprobe A ergab sich für den Faktor AB ein Eigenwert von 2.00 (20 % aufgeklärte Varianz), für den Faktor BS ein Eigenwert von 1,34 (13 % aufgeklärte Varianz) und für Faktor LS ein Eigenwert von 1.09 (11 % aufgeklärte Varianz). Die Kommunalitäten lagen mit Werten Bereich von $0.17 \leq \hat{\boldsymbol{H}} \leq 0.75$ in einem für die Stichprobengröße zufriedenstellendem bis sehr guten Bereich. Auch in Teilstichprobe A wurden mittlere bis hohe Korrelationen zwischen den drei Faktoren gefunden, was erneut bestätigt, dass alle drei Dimensionen inhaltlich eng miteinander verknüpft sind. So korrelieren die Skalen AB und LS mit r = .41, die Faktoren AB und BS mit r = .53 sowie die Faktoren LS und BS mit r = .62.

Für die konfirmatorische Faktorenanalyse in Stichprobe B ergab sich die gleiche Struktur wie in der exploratorischen und konfirmatorischen Faktorenanalyse der Stichprobe A, wiederum mit mittleren bis hohe Korrelationen zwischen den drei Faktoren ($.33 \leq r \leq .53$). Die Annahme des Modells wird hierbei durch gute globale Fit-Indizes gestützt ($\chi^2_{(32)}$ 30.12, p = 0.562; *CFI* ≈ 1.00; *TLI* ≈ 1.00, RMSEA = 0.05; *SRMR* = 0.03).

Tab. 4. Eigenwerte der Faktoren und Anteil aufgeklärter Varianz der resultierenden Skalen für die Faktorenanalyse auf Basis der Daten von Stichprobe B, interne Konsistenz je Skala für A (n = 480) / B (n = 389) / Gesamtstichprobe A+B (N = 869)

Faktor	Interpretation	Items y1-8c	α_C A/B/A+B	Eigenwert	Anteil erklärter Varianz in %	Anteil kumulierter Varianz in %
1	*Verständnisfragen zur Bildentstehung inklusive Abdeckaufgaben (AB)*	y8a, y8b, y8c	.85/.59/.83	2,41	12	12
2	*Geradlinige Lichtausbreitung und Streuung (LS)*	y1, y2, y4, y5	.63/.57/.61	1,45	11	23
3	*Verständnis Bildkonstruktion / Strahlenmodell (BS)*	y3, y6, y7	.54/.43/.45	1,12	10	33

Tab. 5. Antwortverteilung und Ergebnisse der exploratorischen Faktorenanalyse des KTSO-A (Stichprobe B, n = 389)

Variable		Anteil Antworten in %,	Promax Rotation Mustermatrix			
yi	Itemformulierung	Mehrfachauswahl möglich, Gesamtsumme kann > 100 % sein	F1[a]: AB	F2[a]: BS	F3[a]: LS	$\widehat{H}$ (y_i)[b]
y1	***Welche der folgenden Gegenstände / Tiere kann man in einem völlig abgedunkelten Raum sehen?***		-0.09	-0.01	**0.60**	0.33
	ein leuchtendes Glühwürmchen	92 %				
	ein weißes Blatt Papier	7 %				
	einen Fahrrad-Reflektor	6 %				
	die Augen einer Katze	29 %				
y2	***Hat es einen Einfluss auf die Helligkeit in einem Zimmer, ob es helle oder dunkle Tapeten hat?***		0.09	0.13	**0.37**	0.15
	ja, weil die helle Tapete mehr Licht streut, das ins Auge fällt, als eine dunkle Tapete.	77 %				
	nein, weil dunkle Tapeten nichts an der Menge des Lichtes im Raum ändern.	6 %				
	ja, weil auf der hellen Tapete mehr Licht liegen bleibt.	18 %				
	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.	13 %				
y3	***Was ist richtig? Lichtstrahlen sind …***		0.07	**0.22**	0.20	0.14
	…etwas Wirkliches, so wie dünne Wasserstrahlen aus einer Spritzpistole.	19 %				
	…etwas Gedachtes, so wie Konstruktionen in der Geometrie, um z. B. Dreiecks-Probleme lösen zu können.	53 %				
	… exakt das Gleiche wie Lichtbündel.	17 %				
	… Lichtbündel sind etwas Gedachtes, z. B. um die Bildgröße bestimmen zu können.	38 %				
y4	***In einem abgedunkelten Raum ist der helle Fleck einer Taschenlampe an der Wand zu sehen, nicht aber der Lichtstrahl zwischen Taschenlampe und Wand. Warum?***		-0.05	-0.09	**0.49**	0.21
	Erst das an Gegenständen gestreute Licht trifft ins Auge und ist sichtbar.	36 %				
	In dem dunklen Raum wird das Licht absorbiert (verschluckt), daher ist es nicht zu sehen.	29 %				
	Das Licht erhellt die Wand, weil es auf ihr liegen bleibt.	52 %				

Variable		Anteil Antworten in %,	Promax Rotation Mustermatrix			
	Das Licht der Taschenlampe entfernt sich vom Beobachter, erst durch die Wand wird es umgedreht und geht auf den Beobachter zu.	22 %				
y5	***Was passiert, wenn man in dem Lichtstrahl einen Tafellappen aufschüttelt?***		-0.06	0.09	**0.60**	0.44
	Die Staubteilchen wirken wie kleine Linsen, die das Licht auf der Wand bündeln.	7 %				
	Der feine Kreidestaub sammelt das Licht und dadurch sieht man den hellen Fleck auf der Wand nicht mehr.	7 %				
	Die Staubteilchen werden durch das auftreffende Licht durcheinandergewirbelt.	19 %				
	Die Staubteilchen streuen das Licht in alle Richtungen, dadurch trifft es ins Auge und wird sichtbar.	72 %				
y6	***Wie entsteht durch Verwendung einer Sammellinse ein Bild, das auf einem Schirm aufgefangen werden kann?***		0.11	**0.19**	0.13	0.11
	Solch ein Bild entsteht durch Spiegelung der Lichtstrahlen an der Linse nach dem Reflexionsgesetz.	19 %				
	Eine Sammellinse hat den Effekt, die Lichtstrahlen aufzuhellen.	6 %				
	Lichtstrahlen, die von einem Gegenstandspunkt ausgehen, werden durch die Sammellinse abgelenkt und treffen sich im Bildpunkt.	49 %				
	Das Bild geht als Ganzes durch die Linse zum Schirm, dabei wird es in der Linse unter Einhaltung der Linsengesetze umgedreht.	30 %				
y7	***Welche Aussagen zur Bildkonstruktion und Bildentstehung treffen zu?***		-0.08	**0.96**	0.02	0.86
	Nur die ausgezeichneten Strahlen kann man im Strahlengang zeichnen.	15 %				
	Mit den ausgezeichneten Strahlen kann man den Strahlengang besonders leicht zeichnen.	47 %				
	Die ausgezeichneten Strahlen erschweren die Zeichnung, machen sie dafür aber besonders genau.	19 %				
	Ohne die ausgezeichneten Strahlen (wenn diese z. B. durch dünne Stifte aufgehalten werden) kann es kein Bild geben.	20 %				
y8	***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:***					
y8a	***Was passiert, wenn man die untere Hälfte der Linse abgedeckt.***		**0.54**	0.08	-0.15	0.29
	Die obere Hälfte des Bildes wird abgeschnitten.	16 %				
	Die untere Hälfte des Bildes wird abgeschnitten.	20 %				
	Das Bild wird dunkler.	61 %				
	Das Bild wird kleiner.	11 %				
y8b	***Was passiert, wenn man einen Karton mit großem Loch (ringförmige Blende) vor die Linse hält?***		**0.75**	-0.10	0.01	0.51
	Das Bild wird kleiner.	23 %				
	Das Bild wird dunkler.	47 %				
	Die Ränder des Bildes werden kreisförmig abgeschnitten.	37 %				
	Das Bild wird heller.	10 %				
y8c	***Was passiert, wenn man einen Karton mit einem sehr kleinen Loch 5 mm (ringförmige Blende) vor die Linse hält?***		**0.49**	-0.03	0.08	0.26
	Das Bild wird kleiner.	39 %				
	Das Bild wird dunkler.	48 %				
	Die Ränder des Bildes werden kreisförmig abgeschnitten.	27 %				
	Das Bild wird heller.	10 %				

4.3 Validität

Als ein weiterer Baustein der Validität des KTSO-A (neben der Darstellung von Rechercheergebnissen in den Kapiteln 2.2 und 2.3) wurde ein Expertenrating durchgeführt. An dem Rating nahmen 11 Personen (5 Frauen und 6 Männer) im Alter von 27 bis 71 Jahren (M = 40.2 Jahre, SD = 14.8) teil. Alle Befragten verfügten über einen Masterabschluss, Diplom oder Staatsexamen in Physik und hatten sich entweder während des Studiums oder im Anschluss an das Studium auf Fachdidaktik Physik spezialisiert. 8 Befragte haben das Referendariat und das zweite Staatsexamen absolviert und verfügen zwischen 1 bis 30 Jahren Unterrichtserfahrung nach dem Referendariat (M = 13 Jahre, SD = 10.79), 7 Personen aus diesem Kreis haben auch in den Klassenstufen 7 und 8 unterrichtet, eine Person in Klassenstufe 9 und in der Oberstufe. Bei den drei Befragten ohne 2. Staatsexamen handelt es sich um

Doktorand/-innen in Fachdidaktik Physik mit Unterrichtserfahrung an der Hochschule im Umfang von 1 bis 3 Jahren (M = 2.05 Jahre , SD = 1.23). Da sich die Einschätzung der 8 erfahrenen Experten von den weniger erfahrenen Personen deskriptiv kaum und jedenfalls nicht signifikant unterscheiden, werden im Folgenden die Werte für die Gesamtstichprobe aller 11 Befragten berichtet.

Die Befragten stuften jede Testaufgabe des KTSO-A unabhängig voneinander schriftlich (postalische Befragung) und anonym auf Basis der folgenden drei Items anhand einer 5-stufigen endpunktbenannten Likert-Skala von „trifft gar nicht zu (1)“ bis „trifft zu“ (5) ein.

- Item 1: Diese Aufgabe erfasst ein relevantes Konzept der Strahlenoptik.
- Item 2: Ich kann mir vorstellen, diese Aufgabe im Unterricht zur Strahlenoptik in Klassenstufe 7 oder 8 zu thematisieren.
- Item 3: Probleme bei der Lösung dieser Aufgabe zeigen grundlegende Verständnisschwierigkeiten in der Strahlenoptik auf.

In Tabelle 6 wird die unjustierte Intraklassenkorrelation (ICC) für die drei Expertenfragen zu allen 10 Items als Maß der Beurteilerreliabilität sowie das Maß r_{wg} zur Beurteilung des Ausmaßes an Beurteilerübereinstimmung auf Itemebene des KTSO-A berichtet (James, Demaree & Wolf. 1984, S. 87; Lindell, Brandt & Whitney, 1999, S. 127; die Formel, welche der Berechnung zugrunde liegt, ist in Appendix 1 angegeben).

Die unjustierte ICC wurde gewählt, da die Experten die 10 Testaufgaben absolut anhand einer Likert Skala bewerten sollten. Hierbei wurde davon ausgegangen, dass die Beurteiler eine zufällige Auswahl aus der Grundgesamtheit möglicher Beurteiler darstellen. (Die Formel zur Berechnung der verwendeten Intraklassenkorrelation ist ebenfalls in Appendix 1 angegeben). Die ICC ist ein Maß für die Interraterreliabilität einer Ratingskala und kein Maß zur Beurteilung der Übereinstimmung eines einzelnen zu beurteilenden Objekts oder Items.

Der Wert fällt hoch aus, wenn die Beurteiler ein gegebenes Item ähnlich, und verschiedene Items entsprechend unterschiedlich bewerten. Das heißt die Höhe der ICC hängt auch von der Größe der Mittelwertunterschiede zwischen den zu ratenden Items ab.

Insofern ist die Höhe der ICC in diesem Fall unter dem Vorbehalt zu interpretieren, dass idealerweise alle Aufgaben als konsistent geeignet eingestuft werden sollen, was zu geringen Mittelwertunterschieden im Rating der Testaufgaben und damit zu einer geringen Varianz in Bezug auf die Beurteilung führt (Wirtz und Caspar, 2002, S. 161). Vor diesem Hintergrund kann der empirisch vorgefundene Wertebereich von $.50 \leq ICC \leq .62$, wobei alle Werte auf dem 5 % Niveau signifikant sind, als erwartungsgemäß betrachtet werden.

Auf Ebene der Einzelitems des KTSO-A ergeben sich durchweg konsistent Mittelwerte > 4 (Skala 1-5) und bis auf eine Ausnahme (Item 6) verhältnismäßig geringe Standardabweichungen. So wurde Item 6 „Wie entsteht durch Verwendung einer Sammellinse ein Bild, das auf einem Schirm aufgefangen werden kann?“ von einem Teil der Beurteiler kritisch bewertet und Item 8b (ringförmige Abdeckung der Linse) von einem Beurteiler als für diese Altersstufe als schwierig eingestuft. Die konsistenten und bis auf die genannten beiden einzelnen Ausnahmen guten Bewertungen spiegeln sich auch in den Werten der r_{WG} wieder. Im Gegensatz zum ICC, der die Beurteilerreliabilität abschätzt, ist die r_{WG} ein Maß zur Einschätzung der Beurteilerübereinstimmung, welches auf Itemebene angewandt werden kann und gibt an zu welchem Grad unterschiedliche Beurteiler ein Objekt, hier eine Testaufgabe, gleich bewerten (Lindell, Brandt & Whitney. 1999, S. 127). Der Index für Einzelitems vergleicht die beobachtete Varianz der Beurteilerantworten mit der Varianz, welche erwartet wird, wenn sich alle Ratings durch eine gleichverteilte Residualvarianz auszeichnen. Werte $\geq .80$ gelten bei einer fünfstufigen Likert-Skala als ausreichend hoch, um eine konsistente Beurteilerübereinstimmung annehmen zu können (vgl. ebd.). Dieses Kriterium ist für alle Items und jede der drei Expertenfragen mit Ausnahme von Item 6 erfüllt (vgl. Tab. 6).

Als Anhaltspunkt zur Abschätzung der Kriteriumsvalidität wurden des Weiteren Korrelationen zwischen der Gesamtpunktzahl und Zeugnisnoten in Mathematik, Physik und Deutsch berechnet. Im Ergebnis zeigen sich signifikante mittlere Korrelationen zwischen den Fachnoten in Physik ($r_{(869)} = -.34$, $p < 0.001$) und Mathematik ($r_{(869)} = -.29$, $p < 0.001$,) und den Gesamtpunktzahlen des KTSO-A. Die negativen Vorzeichen resultieren aus der entgegengesetzten Skalierung von Schulnoten im deutschen Schulsystem und den Testscores des KTSO-A.

Tab. 8. Expertenrating zum KTSO-A, $N = 11$

	Skala	Item	[a]Expertenbewertung Frage 1 [b]ICC = .522 * (10 Items)			[a]Expertenbewertung Frage 2 [b]ICC = .619 * (10 Items)			[a]Expertenbewertung Frage 3 [b]ICC = .500 * (10 Items)		
			M (SD)	*Median (Range)*	[c]r_{WG}	*M (SD)*	*Median (Range)*	[c]r_{WG}	*M (SD)*	*Median (Range)*	[c]r_{WG}
y_1	*LS*	*Sichtbare Gegenstände / Lebewesen in dunklem Raum*	4.82 (0.40)	5 (4 -5)	0.96	4.73 (0.47)	5 (4 -5)	0.95	4.45 (0.69)	5 (4 -5)	0.89
y_2	*LS*	*Helligkeit eines Zimmers - Farbe der Tapete*	4.64 (0.50)	5 (4 -5)	0.94	4.55 (0.69)	5 (3 -5)	0.89	4.55 (0.69)	5 (3 -5)	0.89
y_3	*BS*	*Lichtstrahlen sind …*	4.45 (0.82)	5 (4 -5)	0.85	4.36 (0.67)	5 (3 -5)	0.90	4.45 (0.82)	5 (4 -5)	0.89
y_4	*LS*	*Lichtstrahl in dunklem Raum*	4.82 (0.40)	5 (4 -5)	0.96	4.82 (0.40)	5 (4 -5)	0.96	4.82 (0.40)	5 (4 -5)	0.95
y_5	*LS*	*Lichtstrahl: Kreidestaub*	4.82 (0.40)	5 (4 -5)	0.96	4.73 (0.47)	5 (4 -5)	0.95	4.82 (0.40)	5 (4 -5)	0.96
y_6	*BS*	*Entstehung reeller Bilder an Sammellinse*	4.10 (1.45)	5 (1 -5)	0.50	4.22 (1.30)	5 (2 -5)	0.62	4.10 (1.20)	5 (2 -5)	0.64
y_7	*BS*	*Bildkonstruktion: Bild-entstehung*	4.60 (0.52)	5 (4 -5)	0.94	4.90 (0.32)	5 (4 -5)	0.98	4.90 (0.32)	5 (4 -5)	0.98
y_{8a}	*AB*	*Abdeckung einer Linsen-hälfte*	4.73 (0.47)	5 (4 -5)	0.95	4.82 (0.40)	5 (4 -5)	0.96	4.91 (0.30)	4 (4 -5)	0.98
y_{8b}	*AB*	*Ringförmige Abdeckung der Linse*	4.67 (0.71)	5 (4 -5)	0.89	4.20 (1.32)	5 (1 -5)	0.88	4.56 (0.73)	5 (3 -5)	0.88
y_{8c}	*AB*	*Enge ringförmige Abde-ckung der Linse*	4.80 (0.42)	5 (4 -5)	0.96	4.50 (0.71)	5 (3 -5)	0.89	4.60 (0.52)	5 (4 -5)	0.94

a 5-stufige Likert Skala von 1 (trifft gar nicht zu) bis 5 (trifft zu)

b Unjustierter Wert, entspricht der Auswahl „absolute agreement“ in SPSS, $*p < 0.05$

c r_{WG} (Definition siehe Appendix 1)

5 DISKUSSION UND AUSBLICK

5.1 Diskussion der Testeigenschaften

Wir diskutieren zunächst die Ergebnisse zu Items, Distraktoren und dem Gesamttest in einem allgemeinen Überblick und dann Grenzen, Mängel und Stärken des Tests im Hinblick auf den Stand der Entwicklung von Konzepttest in anderen Bereichen.

Die Kennwerte des Tests (Itemschwierigkeiten, Trennschärfen, Reliabilität) liegen für Stichprobe A durchwegs und für B großenteils in den in der Literatur angegebenen Akzeptanzbereichen. Abweichungen davon, Unterschiede zwischen den Stichproben (insbesondere die schlechteren Trennschärfen von Stichprobe B im Vergleich zu A) sowie weiterer Entwicklungs- und Verbesserungsbedarf werden unten diskutiert. Die Werte für die Reliabilität (innere Konsistenz) insbesondere sind für beide Stichproben vergleichbar zu denen einer Reihe anderer Konzepttests im naturwissenschaftlich-mathematischen Bereich (Evolution: 0.58, 0.64 - zwei Stichproben), Anderson et al, 2002; Allgemeine Chemie: 0.71, Mulford & Robinson, 2002; elektrische Schaltkreise: 0.7, Engelhardt & Beichner, 2004 ; Elektrizität & Magnetismus: 0.75, Maloney et al., 2001; Statistik: ≈ 0.7 (verschiedene Stichproben), Allen, 2006; Materie-Energie-Wechselwirkung: 0.75, Ding & Beichner, 2009). Der Wert für die Teilstichprobe A ist nahe dem Schwellenwert für Individualdiagnostik (Bühner, 2011; Ding & Beichner, 2009; Doran. 1980).

Die Analyse der Distraktoren zeigt auf, dass alle Distraktoren so plausibel waren, dass sie auch nach dem Unterricht ausreichend häufig gewählt wurden. Niedrige, aber noch ausreichende Werte zwischen 5 % und 10 % ergeben sich bei Items der Skala Lichtausbreitung und Streuung, die vergleichsweise leicht zu lösen waren ($0.67 \leq P_i \leq 0.77$). Dabei zeigt sich, dass auch jeweils zwei der „selteneren“ Distraktoren in gleicher Weise mit einer Häufigkeit über dem Schwellwert gewählt wurden, was als Hinweis gewertet werden kann, dass sie übereinstimmend mit den Ergebnissen von Wiesner (1986; 1992a) tatsächlich relevante Schülervorstellungen erfassen und daher beibehalten werden sollten. Abweichend davon ergibt sich nur bei Item 6, *„Wie entsteht durch die Verwendung einer Sammellinse ein Bild, das auf einem Schirm aufgefangen werden kann?“* ein deutlicher Unterschied in der Wahlhäufigkeit für die Option *„die Linse hat den Effekt, die Lichtstrahlen aufzuhellen“* im Vergleich zu den anderen Distraktoren (6 % versus, 19 % bis 49 %); das Item wurde auch als einziges in dem Expertenrating nicht zufriedenstellend bewertet ($r_{wg} < 0.70$). Dennoch wurde das Item in der Analyse aus inhaltlichen

Gründen einbezogen, da auch dem selten gewählten Distraktor eine klassische Schülervorstellung zugrunde liegt (vgl. dazu Guesne, 1985; Wiesner, 1994), und er selbst nach Optikunterricht noch von immerhin 6 % der Schülerinnen und Schüler gewählt wird.

Es finden sich in den Kennwerten Unterschiede zwischen den Teilstichproben A und B, die sich sinnvoll interpretieren lassen. Die Lösungswahrscheinlichkeit für A ist höher als die für B, was zeigt, dass der Test grundsätzlich in der Lage ist, den Effekt eines auf die fraglichen Lernschwierigkeiten bzw. Fehlkonzepte zielenden Unterrichtes nachzuweisen. Auch die korrigierten Trennschärfen liegen in A generell höher als in B, das gleiche gilt für die interne Konsistenz, ein Sachverhalt, der ähnlich auch bei anderen Konzept-Tests gefunden wurde (Zeilik et al., 1997; Ramlo, 2008). Das lässt sich so deuten, das mit zunehmendem Wissen und Verständnis in einem Bereich diese auch konsistenter werden; Wissen und Verständnis haben Systemcharakter (Mandl & Spada, 1988; Braisby & Gellatly, 2008), insbesondere in den Naturwissenschaften (Nersessian, 1991; Kircher et al., 2009, Kap., 19.4.4).

Das Ergebnis der Faktorenanalyse (Abschnitt 4.2) kann inhaltlich wie folgt interpretiert werden. Es ergeben sich drei Dimensionen des konzeptuellen Verständnisses im Bereich der Strahlenoptik:

- Abdeckaufgaben mit Verständnisfragen zur Bildentstehung (AB);
- Verständnisfragen zur geradlinigen Lichtausbreitung und Streuung (LS) und
- Verständnisfragen zur Bildkonstruktion und zum Strahlenmodell (BS).

Als Anhaltspunkt für die Bewertung der Konstruktvalidität zeigt die Kreuzvalidierung, dass der KTSO-A mit den drei übergeordneten Verständnisbereichen „Abdeckaufgaben mit Verständnisfragen zur Bildentstehung“, „Verständnisfragen zur geradlinigen Lichtausbreitung und Streuung“ sowie „Verständnisfragen zur Bildkonstruktion und zum Strahlenmodell“ das konzeptuelle Verständnis der Strahlenoptik im Kontext der Bildentstehung an der Sammellinse in inhaltlich sinnvoll interpretierbarer Weise erfasst. Das Modell belegt, dass die drei Verständnisbereiche unterschiedliche Facetten des konzeptuellen Verständnisses erfassen, welche jedoch signifikant untereinander korrelieren. Die Korrelationen weisen einen mittleren bis starken Zusammenhang der drei Facetten untereinander auf. Im Zug der Kreuzvalidierung wurde sowohl in der exploratorischen wie auch der konfirmatorischen Faktorenanalyse die gleiche Struktur aufgefunden und zwar, wie zuvor angesprochen, unabhängig davon, ob die Schülerinnen und Schüler gezielt Unterricht erhalten hatte, welcher Schülervorstellungen behandelt (Stichprobe A) oder nicht (Stichprobe B). Die Einschätzung der Kreuzvalidierung und Strukturaufklärung lassen sich so zusammenfassen: 1) die Anwendungskriterien einer exploratorischen Faktorenanalyse zur Strukturaufklärung sind gut erfüllt; 2) sie führt zu einer dreidimensionalen Substruktur des Tests, die gute Fit-Gütekriterien aufweist; 3) die gefundene Struktur wird durch eine konfirmatorische Faktorenanalyse mit ebenfalls guter Fit-Güte bestätigt; 4) die so psychometrisch gefundene Struktur hat eine plausible inhaltliche Deutung; 5) ein Mangel ist, dass 2 von 10 Faktorladungen unter dem üblichen Cut-off-Wert liegen.

5.2 Einordnung der Ergebnisse und Grenzen

Wir diskutieren die gefundenen Ergebnisse nun im Hinblick auf den Stand der Entwicklung von Konzept-Tests in anderen Bereichen. Der Sachverhalt teilweise geringer Item-Test-Korrelationen bei der Entwicklung von Konzepttests ist wohlbekannt (Evolution (CINS): Anderson et al, 2002; Statistik (SCI): Allen, 2006; statische Kräfte (CATS), Dynamik (DCI): Jorion et al., 2015); der in der Literatur diskutierte Grund ist der gleiche, der oben schon bei der Diskussion der Unterschiede von Trennschärfen und inneren Konsistenzen zwischen den Stichproben A und B genannt wurde: Ohne gezielte Intervention für die Konzepte in einem gegebenen Inhaltsbereich (und erst recht überhaupt vor Unterricht) ist die begriffliche Struktur von Lernenden wenig konsistent (zur Konsistenzthematik s. o. und Savinainen & Viiri (2008); Nieminen et al, 2010 und Lasry et al., 2011).

Auch Schwierigkeiten bei der Strukturaufklärung von Konzepttests durch Faktorenanalyse sind in der Literatur wohlbekannt. Bei den wenigsten Tests ist ein Versuch dieser Art publiziert (eine Übersichtsarbeit von, 2012 zählt unter 15 Tests nur 4 Versuche auf; Liu, 2012). Für den CSEM (Elektrizität und Magnetismus, Maloney et al., 2001) ist dieser Versuch gescheitert und hat trotz klarer a priori definierbarer Inhaltsbereiche auf keine interpretierbare Struktur geführt. Für den FCI war und ist eine interpretierbare Faktorstruktur Gegenstand intensiver Auseinandersetzung (Überblick: Lasry et al., 2011), und es hat rund, 20 Jahre gedauert, bis eine solche vorgeschlagen werden konnte (Scott et al., 2012).

Schließlich werden weder für den FCI (Scott et al., 2012) noch für die wenigen anderen Konzepttests, für die heute Faktorstrukturen bekannt sind, Reliabilitäten für die gefundenen Subskalen angegeben (Evolution/CINS: Anderson et al, 2002; elektrische Schaltkreise/DIRECT: Engelhardt & Beicher, 2004; Bewegung, Kraft/FMCE, Ramlo, 2008; Materie-Energie-Wechselwirkung/MIET: Ding & Beichner, 2009). Bei den wenigen Ausnahmen (tatsächlich den einzigen, die den Autoren bekannt sind) bei den auch die Subskalen-Reliabilitäten genannt sind, liegen diese nochmals deutlich unter der Gesamtreliabilität (Allen, 2006, SCI/Statistik: 0.3 – 0.5; Jorion et al., 2015, DCI/Dynamik: 0.2 – 0.6; CATS/statische Kräfte (das beste Beispiel): 0.5. – 0.7). Für den TUG-K, ein anderes gut etabliertes Instrument (kein Konzepttest im eigentlichen Sinne), ergeben sich für die inhaltlich a priori definierten Subskalen 0.18 - 0.57, für die beiden durch PCA gefundenen Hauptfaktoren 0.56 und 0.78 (Bektasli, 2006).

Jorion et al. (2015) geben einen Überblick über die Problematik der Validierung von Konzepttests und illustrieren die oben skizzierten Begrenzungen und Schwierigkeiten bezüglich Validität, Reliabilität und Strukturaufklärung mit einer psychometrischen Analyse von Tests aus drei Bereichen (s.o.). Als Ergebnis ihrer Analyse schlagen sie ein an Zielsetzung und Entwicklungsstand des Forschungsgebietes angepasstes Raster der Bewertung von Konzepttests vor, welches die Kriterien der

klassischen Itemanalyse, der Item-Response-Theorie und der exploratorischen und konfirmatorischen Faktorenanalyse zusammenfasst (Tab. 7). Für den KTSO-A befinden sich alle Werte in dem Bereich zwischen „average“ und „excellent“.

Vor diesem Hintergrund fassen wir die Bewertung der Testgüte des KTSO-A wie folgt zusammen: Es wurden (anhand zweier Stichproben mit Stichprobengrößen $\gtrsim 400$) psychometrische Eigenschaften ermittelt, die – vor allem im Vergleich mit dem Entwicklungsstand von Konzepttests im Allgemeinen – als akzeptabel bis gut eingeschätzt werden können; Unterschiede zwischen verschiedenen Items und zwischen den Stichproben lassen sich sinnvoll und unter Bezug auf die vorliegende Forschung interpretieren.

Im Rahmen der Empfehlungen für Testkonstruktion im Allgemeinen (Haladyna & Downing, 1989) und Konzepttest im Besonderen (Lindell et al., 2007) sehen wir es daher als einen Beitrag der Arbeit an, den Schritt von einer forschungsbasierten Übersicht von inhaltlich relevanten konzeptionellen Schwierigkeiten (2.2) zu einem Instrument mit akzeptabler bis guter psychometrischer Charakterisierung (4.1) vollzogen zu haben.

Tab. 7. Raster zur Bewertung von Konzepttests (“categorical judgement scheme for evaluating a concept instrument” nach Jorion et al, 2015; wir behalten die Bezeichungen des Originals bei), sowie Einordnung des KTSO-A (Rahmen)

Analysis Method	criterion	excellent	good	average	poor	not acceptable
Classical test theory, item statistics	difficulty	**.2** to **.8**[A]	.2 to .8	**.1** to **.9**[B]	1 to .9	0 to .10
	discrimination	**>.2**[A]	>**.1**[B]	>.0	> -.2	> -1.0
	α_C with-item-deleted ≤ overall α_C	**all**	(3)	(6)	(9)	> (9)
	α_C, total score	>.9	>.8	**>.65**[A,B]	>.5	>.0
Item response theory: not applicable, see text						
*Structural analysis**	EFA: items conform to interpretable constructs	**all**	(5)	(10)	(15)	> (15)
	CFA					
	item loadings	>.3	**>.3 (3)**	>.1	>.1 (3)	> -1.0
	CFI	**>.9**	>.8	>.7	>.6	>.0
	RMSEA	<.03	**<.05**	<.10	<.20	>.20

Hinweise:
– Werte in Klammern stehen für Anzahl von Items die das jeweiligen Kriterium verletzen dürfen
– Die Einstufung des KTSO-A in diesem Raster ist in Fettdruck kenntlich gemacht, die Indizes A/B beziehen sich auf die beiden Validierungsstichproben; alle Werte befinden sich in dem Bereich zw. „average“ und „excellent“, s. Rahmen
* Ergebnisse mit EFA zu Stichprobe B, CFA zu Stichprobe A; zusätzlich wurde eine Kreuzvalidierung mit positivem Ergebnis durchgeführt, s. Text.

Darüber hinaus wurde eine mehrdimensionale Struktur gefunden, mit ihrerseits akzeptablen bis z. T. sehr guten Kenngrößen, und einer inhaltlichen Deutung; mit Blick auf die bekannten Schwierigkeiten der Strukturaufklärung bei Konzepttests kann man mit der gebotenen Vorsicht zumindest von starken Hinweisen auf interpretierbare Teilkonstrukte sprechen. In diesem Zusammenhang sei darauf hingewiesen, dass hinsichtlich der üblichen Akzeptanzwerte psychometrischer Indizes durchwegs betont wird, dass diese nicht starre Kriterien darstellen, sondern mit Augenmaß und im Zusammenhang mit dem Stand und Zielsetzungen des Anwendungsbereichs zu sehen sind (Abell et al., 2009; EFPA, 2013; Jorion et al., 2015). Moosbrugger und Kelava (2012) stellen fest, dass „wenn es keine besser geeigneten Testverfahren gibt, der Einsatz eines niedrig reliablen Messinstruments immer noch aufschlussreicher sein kann als der gänzliche Verzicht auf den Einsatz von Tests“. Für den FCI als am intensivsten untersuchten Konzepttest waren Validität, Reliabilität und Strukturaufklärung Gegenstand einer jahrzehntelangen, intensiven wissenschaftlichen Auseinandersetzung (Härtig, 2014; Scott, 2012; Stewart, Griffin & Stewart, 2007; Morris, Branummartin & Harshman et al., 2006; Rebello & Zollman, 2004; Lasry et al., 2011, auch für einen Überblick über ältere Arbeiten). Diese wichtige Diskussion hätte nicht geführt werden können, ohne dass es den Test in publizierter Form gegeben hätte - mit einer mit Augenmaß geführten Bewertung der Gütekriterien. Für den Bereich des KTSO-A liegen ebenfalls seit Jahrzehnten isolierte Items vor, aber kein auch nur in Ansätzen validiertes Instrument. Diese Lücke wird in der vorliegenden Arbeit geschlossen, mit einem nach der oben geführten Diskussion durchaus brauchbaren Instrument, das in sinnvoller Weise Ausgangspunkt einer weiteren Entwicklung in der Community sein kann (wie es etwa beim FCI der Fall war).

Künftige Weiterentwicklungen des KTSO-A in diesem Sinne sehen wir entlang folgender Linien. Auch wenn die Kennwerte (bis auf einzelne Ausnahmen) in den laut Literatur akzeptierten Wertebereichen liegen, müssen wir folgende Einschränkungen einräumen: Die Werte einiger Trennschärfen sind im unteren Bereich des Akzeptanzbereiches; auch die Reliabilität des Gesamttests für die Stichprobe ohne gezielte Intervention für konzeptuelles Lernen ist nur akzeptabel. Hier können Versuche der Überarbeitung der Itemformulierungen und Hinzunahme weitere Items eine Verbesserung bringen; andererseits kann es auch sein, dass wie bei anderen Konzepttest (s. o. g. Diskussion und Literatur) dass die geringe Konsistenz des konzeptionellen Wissens der Lernenden selbst die Ursache ist. Auf der Ebene von Einzelitems können des Weiteren einige Distraktoren optimiert werden: So könnte eine künftige Testfassung die beiden am häufigsten gewählten Distraktoren und die korrekte Antwortoption beibehalten werden (vgl. Rodriguez, 2005, zur optimalen Anzahl an Distraktoren in Multiple-Choice-Tests).

Auch wenn Kreuzvalidierung und inhaltliche Diskussion gute Hinweise auf die Existenz sinnvoll interpretierbarer Subskalen geben, sollten die Reliabilitätswerte der LS und insbesondere BS-Skalen verbessert werden (beispielsweise kann letztere um zwei bis drei weitere Aufgaben erweitert werden, welche speziell das Konzept der Punkt-zu-Punkt-Abbildung thematisieren). Weitere Forschung zu diesem und zu anderen Konzepttests muss aber auch auf der allgemeinen Ebene zeigen, wie die bekannten Probleme der psychometrischen Identifikation und niedriger Reliabilitäten (Maloney et al., 2001; Allen, 2006; Jorion et al., 2015) auch inhaltlich sehr gut definierter Teilskalen zu erklären und ggf. zu lösen sind.

Im Hinblick auf die abgedeckten Teilbereiche bietet es sich im Bereich der Strahlenoptik an, die Entwicklung auf weitere Themen wie Licht- und Schatten, gerichteten Reflexion und Spiegelbildern und virtuellen Bilder bei Spiegeln und Linsen zu erweitern. Dies könnte dazu beitragen, die Lücke zwischen grundsätzlich guter Kenntnis einschlägiger Fehlkonzepte und Lernschwierigkeiten und dem Vorliegen guter Einzelitems hierzu (Wiesner, 1986, 1992a, 1992b; Fetherstonhaugh & Treagust, 1992; Chu et al., 2009) und psychometrisch validierten Konzepttests zu schließen. Eine Weiterentwicklung des vorliegenden Tests in diesem Sinn ist in Arbeit.

Schließlich ist eine Begrenzung der vorliegenden Arbeit, dass die Validierung bei einer mehrheitlich aus dem Gymnasium stammenden Stichprobe durchgeführt wurde. Bei anderen Schülergruppen (z. B. in der Realschule oder integrierten Gesamtschule), kann es zu Abweichungen in den Ergebnissen zu den Itemkennwerten, der internen Konsistenz oder der Dimensionalität kommen. Ob Schülervorstellungen gezielt im Unterricht behandelt wurden (Teilstichprobe A) oder nicht (Teilstichprobe B) wirkt sich aber, wie die Kreuzvalidierung belegt, nachweislich nicht auf die Struktur des Tests aus.

5.3 Ausblick und Fazit

Für die o. g. Nachteile und Grenzen des Tests können nur weitere Untersuchungen zu dessen Einsatz Aufschluss geben, bei verschiedenen Lernergruppen, und durch verschiedene Forschergruppen, wie es auch die Entwicklung des FCI über mehr als zwei Jahrzehnte vorangebracht hat. Dazu erscheint es sinnvoll, den Test in der heutigen Fassung in publizierter Form zur Verfügung zu haben. Bereits auf dem gegenwärtigen Stand kann als Resümee festgehalten werden, dass der KTSO-A mit den vorgestellten 10 Items eine laut Literaturstandards psychometrisch akzeptable bis gute und zugleich praktikable Testfassung für Konzepte zur Bildentstehung an der Sammellinse und der geradlinigen Lichtausbreitung und Streuung darstellt, die für Forschungszwecke und zur unterrichtlichen Diagnostik eingesetzt werden kann.

APPENDIX

Appendix 1:

ICC $_{\text{unjustiert / random}}$ (Wirtz & Caspar, 2002, S. 184)

$$ICC_{unjust} = \frac{MS_{zw} - MS_{res}}{MS_{zw} + (k-1)\,MS_{res} + \frac{k}{n}(MS_{rat} - MS_{res})}$$

(MS = „mean squares“)

MS_{zw}: Varianz zwischen den zu ratenden Objekten

MS_{res}: Residualvarianz

MS_{rat}: Varianz zwischen den Beurteilern

k: Anzahl der Beurteiler (hier 11)

n: Anzahl der zu beurteilenden Objekte oder Personen (hier 10 Testitems)

r_{wg} Index of Interrater Agreement for Multi-Item Ratings of a single target (James, Demaree, Wolf. 1984, S. 87; Lindell, Brandt & Whitney, 1999, S. 127)

$$r_{wg} = 1 - \frac{s_x^2}{s_{EU}^2}$$

r_{wg}: Beurteilerübereinstimmung innerhalb einer Gruppe von k Beurteilern in Bezug auf ein einzelnes zu beurteilendes Objekt X

s_x^2: beobachtete Varianz in Bezug auf das beurteilte Objekt X

s_{EU}^2: Varianz in Bezug auf das beurteilte Objekt X, die zu erwarten wäre, wenn alle Urteile nur auf einem zufälligen Messfehler basieren.
Der Index „EU“ steht für einen erwarteten Fehler E basierend auf einer Gleichverteilung englisch „uniform distribution“ mit A Antwortmöglichkeiten;
bei einer fünf-stufigen Likertskala ergibt sich: $s_{EU}^2 = \frac{A^2-1}{12} = 2$

Appendix 2: Vollständige Testfassung mit Lösungen

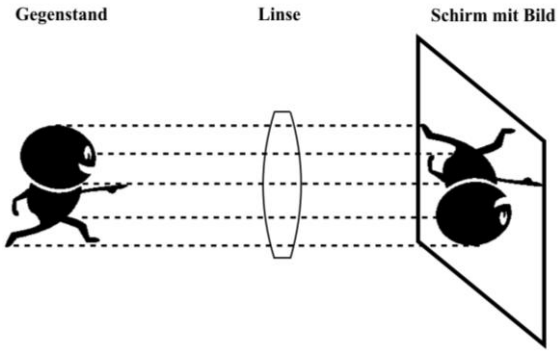
Konzepttest Strahlenoptik Abbildungen

Kreuze jeweils diejenigen Antworten an, die richtig sind.
Es können eine oder mehrere Antworten richtig sein! Beachte die Abbildungen, wenn vorhanden!

Skala	Nr.	Item
LS	1.	**Welche der folgenden Gegenstände / Lebewesen kann man in einem völlig abgedunkelten Raum sehen?** ❑ ein leuchtendes Glühwürmchen ❑ ein weißes Blatt Papier ❑ einen Fahrrad-Reflektor ❑ die Augen einer Katze
LS	2.	**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. ❑ 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.
BS	3.	**Was ist richtig?** ❑ Lichtstrahlen sind etwas Wirkliches, so wie dünne Wasserstrahlen aus einer Spritzpistole. ❑ Lichtstrahlen sind etwas Gedachtes, so wie Konstruktionen in der Geometrie, um z. B. Dreiecks-Probleme lösen zu können. ❑ Lichtstrahlen sind exakt das gleiche wie Lichtbündel. ❑ Lichtbündel sind etwas Gedachtes, z.B. um die Bildgröße bestimmen zu können.
LS	4.	**In einem abgedunkelten Raum ist der Lichtfleck einer Taschenlampe an der Wand zu sehen, nicht aber der Lichtstrahl von der Taschenlampe zur Wand. Warum?** ❑ Erst das an Gegenständen gestreute Licht trifft ins Auge und ist sichtbar. ❑ 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 die Wand wird es umgedreht und geht auf den Beobachter zu.
LS	5.	**Was passiert, wenn man in dem Lichtstrahl einen Tafellappen aufschüttelt?** ❑ Die Staubteilchen wirken wie kleine Linsen, die das Licht auf der Wand bündeln. ❑ 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.

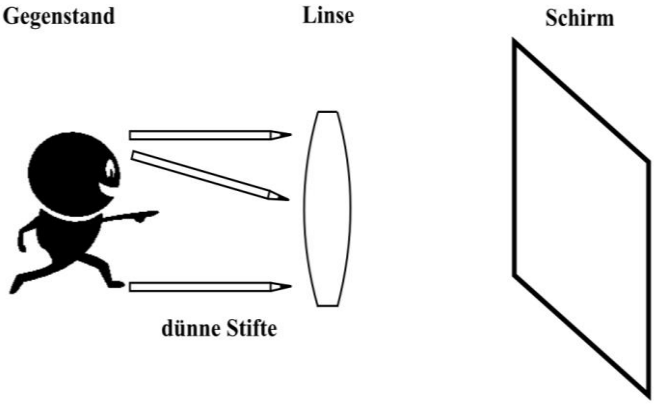
6. **Wie entsteht durch Verwendung einer Sammellinse ein Bild, das auf einem Schirm aufgefangen werden kann?** [1]

- ❑ Solch ein Bild entsteht durch Spiegelung der Lichtstrahlen an der Linse nach dem Reflexionsgesetz.
- ❑ Eine Sammellinse hat den Effekt, die Lichtstrahlen aufzuhellen.
- ❑ 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 (siehe Skizze).



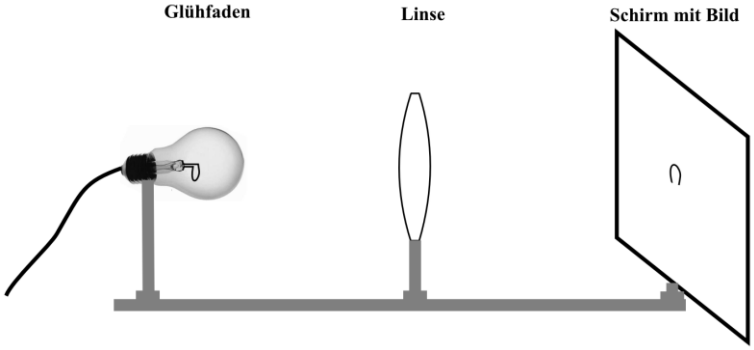
BS 7. **Welche Aussagen zur Bildkonstruktion und Bildentstehung treffen zu?** [2]

- ❑ Nur die ausgezeichneten Strahlen kann man im Strahlengang zeichnen.
- ❑ Mit den ausgezeichneten Strahlen kann man den Strahlengang besonders leicht zeichnen.
- ❑ 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 (siehe Abbildung).

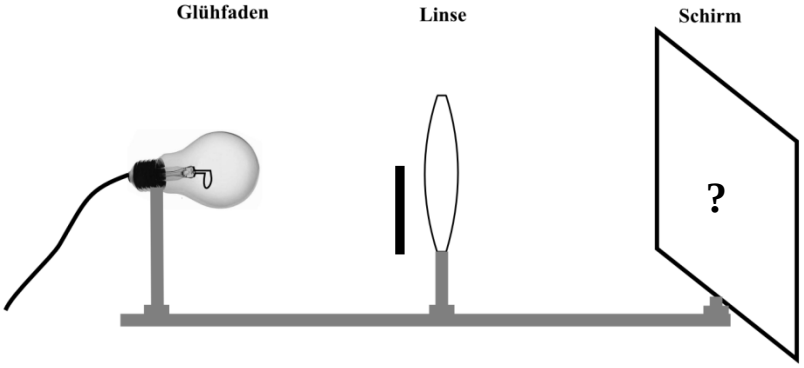


[1] Abbildungen zu Item 6 in Anlehnung an Wiesner, 1986, S. 28
[2] Abbildungen zu Item 7 in Anlehnung an Wiesner, 1986, S. 28

8. **In einer Versuchsanordnung sind eine Glühlampe, eine Sammellinse und ein Schirm so montiert, dass ein vergrößertes, umgekehrtes, scharfes Bild des Glühfadens entsteht:**[3]



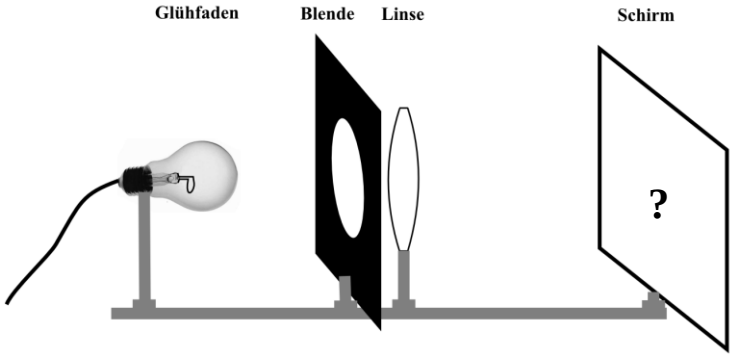
a) **Was passiert, wenn die untere Hälfte der Linse abgedeckt wird?**[4]



AB

- ❑ Die obere Hälfte des Bildes wird abgeschnitten.
- ❑ Die untere Hälfte des Bildes wird abgeschnitten.
- ❑ Das Bild wird dunkler.
- ❑ Das Bild wird kleiner.

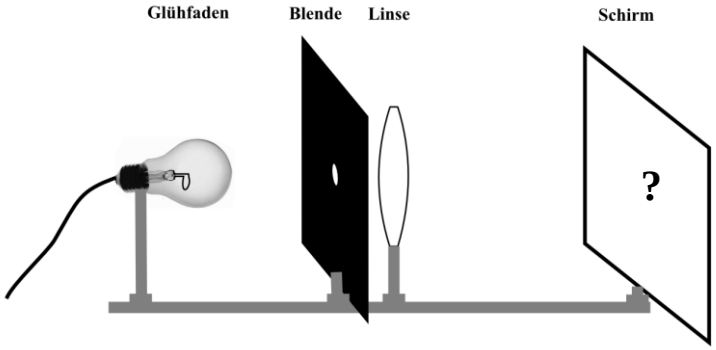
b) **Was passiert, wenn man einen Karton mit großem Loch (ringförmige Blende) vor die Linse hält?** [5]



- ❑ Das Bild wird kleiner.
- ❑ Das Bild wird dunkler.
- ❑ Die Ränder des Bildes werden kreisförmig abgeschnitten.
- ❑ Das Bild wird heller.

[3] Abbildung zu Item 8a-c in Anlehnung an Goldberg und McDermott, 1987, S. 112
[4] Abbildung zu Item 8a in Anlehnung an Goldberg und McDermott, 1987, S. 112
[5] Abbildung zu Item 8b Anlehnung an Goldberg und McDermott, 1987, S. 112

c) **Was passiert, wenn man einen Karton mit einem sehr kleinen Loch (5 mm ringförmige Blende) vor die Linse hält?** [6]



- ❑ Das Bild wird kleiner.
- ❑ Das Bild wird dunkler.
- ❑ Die Ränder des Bildes werden kreisförmig abgeschnitten.
- ❑ Das Bild wird heller.

Lösungen:

Die Lösungen sind wie folgt angeben
(Beispiel, hier Item 1)

Welche der folgenden Gegenstände / Lebewesen kann man in einem völlig sehen abgedunkelten Raum?

- x ein leuchtendes Glühwürmchen
- ❑ ein weißes Blatt Papier
- ❑ einen Fahrrad-Reflektor
- ❑ die Augen einer Katze

Lösung: Nr. 1: a)

Tabelle der Lösungen

Item	Antwort
1	a)
2	a)
3	b)
4	a)
5	d)
6	c)
7	b)
8a	c)
8b	b)
8c	b)

REFERENZEN

Abell, N., Springer, D.W. & Kamata, A. (2009). Developing and validating rapid assessment instruments. Oxford: University Press.

Adams, W. K. & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *Int. J. Sci. Educ.* 33, 1289-1312.

Allen, K.C. (2006). The Statistic Concept Inventory: Development and Analysis of a Cognitve Assessment Instrument in Statistics, Doctoral Dissertation. Oklahoma: University of Oklahoma.

Anderson, D.L., Fisher, K.M. & Norman, G.J. (2002). Development and Evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*. *39*(10), 952-978.

Andersson, B. & Kärrqvist, C. (1983). How Swedish pupils, aged 12-15 years, understand light and its properties. *European Journal of Science Education*, *5*(4), 387-402.

Bardar, E. M., Prather, E. E; Brecher, K. & Slater, T. F. (2007). Development and validation of the light and spectroscopy concept inventory. *Astronomy Education Review*, *5*(2), 103-113.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, &, Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2001). PISA, 2000: Schülerleistungen im internationalen Vergleich.

[6] Abbildung zu Item 8c in Anlehnung an Goldberg und McDermott, 1987, S. 112

Bektasli, B. (2006). The relationships between spatial ability, logical thinking, mathematics performance and kinematics graph interpretation skills of 12th grade physics students. Doctoral Dissertation. Columbus: Ohio State University.

Bonanomi, A., Nai Ruscone, M. & Osmetti, S. A. (2013). The Polychoric Ordinal Alpha, measuring the reliability of a set of polytomous ordinal items. SIS, 2013 Statistical Conference. *Vita e Pensiero*. 1-6.

Braisby, N. & Gellatly, A. (2008). Cognitive Psychology Oxford: University Press.

Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion (3. Aufl.). München, Boston [u.a.]: Pearson Studium.

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology*, *27*(5), 777-786.

Chu, H. E., Treagust, D. & Chandrasegaran, A. L. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science and Technological Education*, *27*(3), 253-265.

Ding, L. & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research*, *5*(2), 020103.

Doran, R. L. (1980). Basic Measurement and Evaluation of Science Instruction. Washington, DC: National Science Teachers Association.

Duit, R. & Treagust, D.R. (2003). Conceptual change: a powerful framework for improving science teaching and learning, *International Journal of Science Education*, *25*(6), 671-688.

Duit, R. (2009). Bibliography STCSE. Students and teachers conceptions and science education. http://archiv.ipn.uni-kiel.de/stcse/: Datum des letzten Aufrufs 13.04.2020.

Engelhardt, P. V. & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics*, *72*(1), 98-115.

EFPA (2013). EFPA Review Model for the description and evaluation of psychological and educational tests. Brussels: European Federation of Psychology Associations.

Eid, M., Gollwitzer, M. & Schmitt, M. (2011). Statistik und Forschungsmethoden. Lehrbuch; mit Onlinematerialien (2. Aufl.). Weinheim [u.a.]: Beltz.

Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D. & Borsboom, D. (2012) qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48 (4), 1 18. http://www. jstatsoft.org/v48/i04/: Datum des letzten Aufrufs 15.04.2020.

Evers, A. (2001). Improving Test Quality in the Netherlands: Results of 18 years of Test Ratings. *International Journal of Testing*, 1, 137-153.

Fetherstonhaugh, T. & Treagust, D. F. (1992). Students' understanding of light and its properties: Teaching to engender conceptual change. *Science Education*, *76*(6), 653-672.

Fox, J. (2010). Package 'polycor': Polychoric and Polyserial Correlations. Version 0.7-8. Freie Statistiksoftware R. http://cran.r-project.org/web/ packages/polycor/: Datum des letzten Aufrufs 02.02.2012. Aktualisierte Version (2019) unter: https://cran.r-project.org/web/packages/polycor/polycor.pdf. Datum des letzten Aufrufs 02.04.2020.

Galili, I. (1996). Students' conceptual change in geometrical optics. *International Journal of Science Education. 18*(7), S. 847-868.

Gerdes, J. & Schecker, H. (1999). Der Force Concept Inventory - Ein diagnostischer Test zu Schülervorstellungen in der Mechanik. *Der Mathematische und Naturwissenschaftliche Unterricht, 52*(5), S. 283-288.

Goldberg, F. M. & McDermott, L. C. (1987). An investigation of student understanding of the real image formed by a converging lens or concave mirror. *American Journal of Physics*, *55*(2), 108-119.

Guesne, E. (1985). Light. R. Driver, E. Guesne und A. Tiberghien (Hrsg.), Children's ideas in science (S. 10–32). Milton Keynes [Buckinghamshire], Philadelphia: Open University Press.

Guzzetti, B. J., Snyder, T. E., Glass, G. V. & Gamas, W. S. (1993). Promoting Conceptual Change in Science: A Comparative Meta-Analysis of Instructional Interventions from Reading Education and Science Education. *Reading Research Quarterly*, *28*(2), 117-155.

Haagen-Schützenhöfer, C. & Hopf, M. (2012). Entwicklung eines Testinstruments zur geo-metrischen Optik. In Bernholt, S. (Hrsg.), GDCP Jahrestagung in Oldenburg, 2011: Konzepte fachdidaktischer Strukturierung für den Unterricht. Münster/New York: LIT Verlag.

Haagen-Schützenhöfer, C. & Hopf, M. (2014a). Development of a two-tier test-instrument for geometrical optics. In: C. P. Constantinou, N. Papadouris und A. Hadjigeorgiou (Eds.), *E-Book Proceedings of the ESERA, 2013 Conference: Science Education Research For Evidence-based Teaching and Coherence in Learning*. (pp. 24-30) (Nicosia/Cyprus: European Science Education Research Association., 2014).

Haagen-Schützenhöfer, C. & Hopf, M. (2014b). Testing students' conceptual understanding in geometrical optics with a two-tier instrument. In Taşar, M. (Hrsg.), World Conference of Physics Education, 2012 in Istanbul: *Book of Proceedings* (Ankara: Pegem Akademi, 2014); (pp. 1327-1336).

Haagen-Schützenhöfer, C. (2014c). Students' conceptions on the nature of white light. In Dvorak, L., Koudelková, V. (Hrsg.), ICPE-EPEC Conference, 2013: Active learning – in a changing world of new technologies (MATFYZPRESS publisher, Prague, 2014); http://iupap-icpe.org/publications/proceedings/ICPE-EPEC_2013 _proceedings.pdf: Datum des letzten Aufrufs 29.09.2014.

Haagen-Schützenhofer, C. & Hopf, M. (2018) Schülervorstellungen zur geometrischen Opti. (S. 89 – 114). In: H. Schecker,, T. Wilhelm, M. Hopf, M.& R. Duit. Schülervorstellungen und Physikunterricht: Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis. Springer Spektrum: Berlin, Heidelberg.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 1, 51-78.

Härtig, H. (2014). Der Force Concept Inventory Vergleich einer offenen und einer geschlossenen Version. *PhyDid A-Physik und Didaktik in Schule und Hochschule, 1*(13), 53-61.

Herdt, D. (1990). Einführung in die elementare Optik. Vergleichende Untersuchung eines neuen Lehrgangs. Essen: Westarp-Wissenschaftsverlag.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force concept inventory. *Phys. Teach*, *30*(3), 141-158.

Hettmannsperger, R. (2015). Lernen mit multiplen Repräsentationen aus Experimenten: Ein Beitrag zum Verstehen physikalischer Konzepte. Springer Verlag Sozialwissenschaften: Wiesbaden.

Heywood, D. S. (2005). Primary trainee teachers' learning and teaching about light: Some pedagogic implications for initial teacher training. *International Journal of Science Teaching, 27*(12). 1447-75.

Horn, J. (1965). A rationale and test fort the number of factors in factor analysis. *Psychometrika* (30), 179-185.

James, L. R., Demaree, R. G. & Wolf, G. (1984). Estimating withingroup interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85-98.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*(4), 454-496.

Jung, W. (1981). Ergebnisse einer Optik-Erhebung. *Physica Didactica* 9. 19-34.

Kircher, E.; Girwidz, R.; Häußler, P. (JAHR) Physikdidaktik - Theorie und Praxis, Springer Verlag, Berlin, Heidelberg, New York.

Kline, Th. J. B. (2005). Psychological Testing. A Practical Approach to Design and Evaluation. Thousand Oaks, London, New Delhi: Sage.

Kline, P. (2015). A Handbook of Test Construction: Introduction to psychometric design London: Routledge.

Langley, D., Ronen, M. & Eylon, B.-S. (1997). Light propagation and visual patterns: Preinstruction learners' conceptions. *Journal of Research in Science Teaching*, *34*(4), 399-424.

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A. & Reshef, O. (2011). The puzzling reliability of the Force Concept Inventory. *American Journal of Physics*, *79*(9), 909-912.

Lienert, G. A. & Raatz, U. (1998). Testaufbau und Testanalyse. 6. Aufl. Weinheim: Beltz. Psychologie Verl.-Union.

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: a critical appraisal. *Learning and Instruction*, 11, (4-5). 357-380.

Lindell, M. K., Brandt, Ch. J. & Whitney, D. J. (1999). A Revised Index of Interrater Agreement for a Single Target. *Applied Psychological Measurement*, *23*(2), 127-135.

Lindell, R. S., Peak, E. & Foster, Th.M. (2007). Are They All Created Equal? A Comparison of Different Concept Inventory Development Methodologies., 2006 *Physics education research conference. AIP Conference Proceedings*, Volume 883, 14-17.

Liu, X. (2012). Developing Measurement Instruments for Science Education Research. Second International Handbook of Science Education. (Springer International Handbooks of Education). Barry J. Fraser, Kenneth Tobin, Campbell J. McRobbie (Eds.). Springer. Dortrecht. 651-665.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2017). Resource Letter RBAI-1: Research-Based Assessment Instruments in Physics and Astronomy. *American Journal of Physics*, *85*(4), 245-264.

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J. & van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *Am. J. Phys. Phys. Educ. Res., Suppl*. 69, 12-23.

Mandl. H. & Spada, H. (Hrsg.). (1988): Wissenspsychologie. Psychologie Verlags-Union, München.

Martinez-Borreguero, G., Pérez-Rodríguez, Á. L., Suero-López, M. I., & Pardo-Fernández, P. J. (2013). Detection of misconceptions about colour and an experimentally tested proposal to combat them. *International Journal of Science Education*, *35*(8), 1299-1324.

Meschede, D. (2006) (Hrsg.) Gerthsen Physik. (23. Aufl.) Berlin: Springer.

Moosbrugger, H., & Kelava, A. (2012). Testtheorie und Fragebogenkonstruktion. Berlin: Springer.

Morris, G. A., Branummartin, L., Harshman, N. et al. (2006). Testing the test: item reponse curves and test quality. *American Journal of Physics*, 74, 449-453.

Mulford, D. R. & Robinson, W. R. (2002). An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Educ*., *79*(6), 739-744.

Naranjo-Correa, F. L., Martinez-Borreguero, G., Perez-Rodriguez, A. L., Lopez, S., Isabel, M., & Pardo Fernandez, P. J. (2015). A new online tool to detect color misconceptions. Color Research & Application.

Nersessian, N. J. (1992). How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science. In: R. N. Giere (Hrsg.), Cognitive models of science: Vol. 15. *Minnesota studies in the philosophy of science* (pp. 129-186) Minneapolis, MN: University of Minnesota Press.

Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force concept inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics Physics Education Research*, *6*(2), 1-12.

Özdemir, G. & Clark, D. B. (2007). An Overview of Conceptual Change Theories. *Eurasia Journal of Mathematics, Science and Technology Education*, *3*(4), 351-361.

Ramlo, S. (2008). Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics*, *76*(9), 882-886.

Rebello, N. S. & Zollman, D. A. (2004). The effect of distracters on student performance on the force concept inventory. *American Journal of Physics*, *72*(1), 116-125.

Reiner, M., Slotta, J. D., Chi, M. T. H. & Resnick, L. B. (2000). Naive Physics Reasoning: A Commitment to Substance-Based Conceptions. *Cognition and Instruction*, *18*(1), 1-34.

Revelle, W. (2013) psych [Computer Software]: freie Statistiksoftware R. http://cran.r-project.org/web/ packages/psych/index.html: Datum des letzten Aufrufs 22.09.2013. Neueste Version: Revelle W (2019). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 1.9.12, https://CRAN.R-project.org/package=psych. Datum des letzten Aufrufs 17.04.2020.

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, *24*(2) 3-13.

Rosseel, Y. (2012). Package lavaan [Computer software]: freie Statistiksoftware R. https://users.ugent.be/~yrosseel/lavaan/lavaanIntroduction.pdf. Datum des letzten Aufrufs 18.04.2020.

Rost, J. (2004). Lehrbuch Testtheorie - Testkonstruktion. 2. Auflage Bern, Göttingen [u.a]: Huber.

Savinainen, A., & Viiri, J. (2008). The Force Concept Inventory as a measure of students conceptual coherence. International *Journal of Science and Mathematics Education*, *6*(4), 719-740.

Schecker, H. (o.J.). Geometrische Optik, Testaufgaben. http://www.idn.uni-bremen.de/schuelervorstellungen/. Datum des letzten Aufrufs 02.04.2020.

Scheid, J. (2013). Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung: Ein Beitrag zur Entwicklung der Aufgabenkultur. Dissertationsschrift zur Erlangung des Doktors der Philosophie. In Niedderer, H., Fischler, H., Sumfleht E. (Hrg.). Studien zum Physik- und Chemielernen Band, 151, Berlin: Logos Verlag.

Scheid, J., Müller, A., Hettmannsperger, R. & Schnotz, W. (2019). Improving learners' representational coherence ability with experiment-related representational activity tasks. *Physical review of physics education research* 15, 010142. DOI: 10.1103/PhysRevPhysEducRes.15.010142

Scott, P. H., Adams, H., & Leach, J. (2007). Student Conceptions and Conceptual Learning in Science. In S. K. Abell & N. G. Lederman (Eds.), Handbook of Research on Science Education (pp. 31-54).Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Scott, T. F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics-Physics Education Research*, *8*(2), 020105.

Selley, N. J. (1996). Children's ideas on light and vision. *International Journal of Science Education*, *18*(6), 713-723.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In Barner, D., & Barron, A. S. (Eds.), Core knowledge and conceptual change (pp. 53-72). New York: Oxford University Press.

Sokoloff, D. R. (2006). Action Research and the Light and Optics Conceptual Evaluation. (227 – 243). In: Active learning in optics and photonics: training manual. D.R. Sokoloff (Ed.). ALOP Manuel Training English version, UNESCO 2006. https://unesdoc. unesco.org/ark:/48223/pf0000217100. Datum des letzten Aufrufs 26.04.2020.

Stewart, J., Griffin, H. & Stewart, G. (2007). Context sensitivity in the force concept inventory. *Physics Review Special Topics – Physics Education Research*, 3, 010102.

Strobl, C. (2012). Das Raschmodell. Eine verständliche Einführung für Studium und Praxis. München und Mering: Rainder Hampp.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics. Boston: Allyn and Bacon.

Tyson, L. M., Venville, G. J., Harrison, A. G. & Treagust, D. F. (1997). A Multidimensional Framework for Interpreting Conceptual Change Events in the Classroom. *Science Education*, *81*(4), 387-404.

Wiesner, H. (1986). Schülervorstellungen und Lernschwierigkeiten im Bereich der Optik. *Naturwissenschaften im Unterricht - Physik*, *34*(13), 25-29.

Wiesner, H. (1992a). Schülervorstellungen und Lernschwierigkeiten mit dem Spiegelbild. *Naturwissenschaften im Unterricht - Physik*, *3*(14), 16-18.

Wiesner, H. (1992b). Verbesserung des Lernerfolgs im Unterricht über Optik (1). *Physik in der Schule*, *30*(9), 286-290.

Wiesner, H. (1994). Ein neuer Optikkurs für die Sekundarstufe I, der sich an Lernschwierigkeiten und Schülervorstellungen orientiert. Naturwissenschaften im Unterricht - Physik, *5*(22), 7-15.

White, R. T. & Gunstone, R. F. (1989). Metalearning and conceptual change. *International Journal of Science Education, 11*(5), 577-586.

Wilcox, R. R. (1981). Analyzing the Distractors of Multiple-Choice Test Items or Partitioning Multinomial Cell Probabilities with Respect to a Standard. *Educational and Psychological Measurement*, *41*(4), 1051-1068.

Wirtz, M. & Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. München: Hogrefe.

Yeo, S., & Zadnik, M. (2001). Introductory Thermal Concept Evaluation: Assessing Students' Understanding. *The Physics Teacher*, 39, 495 504. Ins Deutsche übersetzt von Engelke, T. abrufbar unter http://www.didaktik.physik.uni-muenchen.de/forschung/testdatenbank/inhalt_testdatenbank/verst_waermelehre.pdf: Datum des letzten Aufrufs 02.05.2015. (passwortgeschützt).

Zeilik, M., Schau, C., Mattern, N., Hall, S., Teague, K. W. & Bisard, W. (1997). Conceptual astronomy: A novel model for teaching postsecondary science courses. *American Journal of Physics*, *65*(10), 987-996.