CONCEPT MAPS ALS DIAGNOSEINSTRUMENT IM PHYSIKUNTERRICHT UND DEREN AUSWIRKUNG AUF DIE DIAGNOSEGENAUIGKEIT VON PHYSIKLEHRKRÄFTEN

Dissertation
von Siv Ling Ley
aus Gelsenkirchen

eingereicht zur Erlangung
eines Doktorgrades der Naturphilosophie (Dr. phil. nat.)
an der Fakultät für Physik
der Universität Duisburg-Essen

- im Oktober 2014 -

1. Gutachter: Prof. Dr. Hans E. Fischer

2. Gutachter: Prof. Dr. Helmut Fischler

Tag der mündlichen Prüfung: 08. April 2015

Dieses Dissertationsprojekt ist in der DFG-geförderten Forschergruppe und dem Graduiertenkolleg „naturwissenschaftlicher Unterricht-essen, nwu“ der Universität Duisburg-Essen in der Arbeitsgruppe von Prof. Dr. Hans E. Fischer durchgeführt worden. Darüber hinaus förderte die Stiftung Mercator im Rahmen des Projekts „Ganz In - mit Ganztag mehr Zukunft. Das neue Ganztagsgymnasium“ diese Arbeit.
Allen Institutionen vielen Dank für die finanzielle Realisierung dieses Projekts und der damit verbundenen Perspektiven.

Inhaltsverzeichnis

Kurzfassung

Bei deutschen Mathematik- und Deutschlehrkräften sind Defizite bei diagnostischen Kompetenzen zu erkennen. Für Physiklehrkräfte ist die Forschungslage so dürftig, dass Aussagen zu ihrer Diagnosekompetenz momentan nicht getroffen werden können. Zudem gibt es für den Physikunterricht nur wenige verlässliche Diagnoseinstrumente. Ziel dieser Arbeit ist die Entwicklung eines Diagnoseinstrumentes, mit dem Schülerinnen und Schüler zeitnah im Unterricht eingeschätzt werden können. Die Entwicklung des Instruments wird außerdem genutzt, die Diagnosekompetenz der beteiligten Physiklehrkräfte in Form von Diagnosegenauigkeit einzuschätzen. Hierzu wird in zwei Studien ein Verfahren zum Einsatz von Concept Maps mit unterschiedlichen Aufgaben- und Bewertungsformaten entwickelt.

Die Entwicklung, Pilotierung und Validierung des Diagnoseinstruments ‚Concept Map' wird in der ersten Studie durchgeführt. Die Ergebnisse zeigen, dass das entwickelte Concept Map-Aufgabenformat und das Bewertungsformat ‚Concept Map-Beurteilungsbogen' partiell Kompetenzen der Schülerinnen und Schüler abbilden können, wie sie in einem Kompetenztest gemessen werden (r = .29*, p < .05). In Einklang mit anderen Ergebnissen der Concept-Map- Forschung kann von einer konvergenten Validität im unteren Korrelationsbereich gesprochen werden.

Mit den in der ersten Studie entwickelten Instrumenten wird die Diagnosegenauigkeit der Lehrkräfte in einer zweiten Studie als Rangkorrelation gemessen. Mit einer Stichprobe von 48 Physiklehrkräften mit ihren 977 Schülerinnen und Schülern, konnten Gruppenunterschiede hinsichtlich der Diagnosegenauigkeit (H (3) = 10.77, p < .05, ω = .47) festgestellt werden. Lehrkräfte, die ihre Schülerinnen und Schüler anonym anhand einer Concept Map mit Hilfe des Beurteilungsbogens bewerten, können genauso gut eine Rangordnung ihrer Schülerinnen und Schüler bilden, wie Lehrkräfte, die ihre Schülerinnen und Schüler personalisiert auf Basis ihrer Unterrichtsbeobachtungen einschätzen. Zusammenfassend ermöglicht das entwickelte Concept Map Verfahren mit Bewertungsbögen eine anonyme Beurteilung der Schülerfähigkeiten mit einer Diagnosegenauigkeit, die ähnlich erfolgreich ist wie die Beurteilung, die die Kenntnis der Schülerfähigkeiten über einen längeren Unterrichtsabschnitt voraussetzt. Es kann außerdem erwartet werden, dass sich die Genauigkeit durch eine entsprechende Ausbildung der Lehrpersonen steigern lässt.

Abstract

The empirical research showed: German teachers of the subjects German and Mathematics do not have optimal diagnostic competences measured as diagnostic accuracy. The current state of research especially for physics teachers is quite low. Up to now, statements about physics teachers' diagnostic competence cannot be made clearly. In addition, for the subject physics exists a lack of diagnostic instruments. The aim of this work is to develop a diagnostic instrument which allows for students' diagnostic in lessons. The diagnostic instrument will also be used to measure teachers' diagnostic accuracy in form of diagnostic accuracy. In two studies concept maps with different task formats and scoring formats will be developed.

The subject of the first study is the development, piloting and validation of the diagnostic instrument 'concept map'. The results show that the concept map-task format, developed in this study, and the concept map-scoring format 'concept map-evaluation sheet' measure competences partially as they can be measured in a competence test (r = .29*, p < .05). Based on the results and in accordance to already existing research results about concept maps, a convergent validity on a lower level can be assumed.

Physics teachers' diagnostic instrument will be measured as a rank correlation with the developed instruments of study 1. The analysis of 48 physics teachers, who participated with 977 students, shows a general group difference regarding diagnostic accuracy (H (3) = 10.77, p < .05, ω = .47). Teachers who assess their students anonymously using a concept map and the evaluation sheet, are as well successful in performing a rank order as teachers who assess their students personalized by using their previous observations and experiences. In summary, concept maps with the evaluation sheet allow a more objective students' assessment in comparison to a judgement which assumes the knowledge of students' abilities over a long term. It can be expected that the accuracy can be enhanced by appropriate training of teachers.

1 Einleitung

Seit etwa dem Jahr 2000 hat sich in deutschen Schulen eine neue Sicht auf Lehr-Lern-Prozesse durchgesetzt. Statt Wissen zu erwerben, sollen Schülerinnen und Schüler Kompetenzen fachspezifisch und fächerübergreifend aufbauen. Damit sind ebenfalls neue Anforderungen an die Curricula der lehrerausbildenden Institutionen entstanden. Angehende Lehrerinnen und Lehrer an der Hochschule und Absolventen im Anfangsschuldienst sollen kompetenzorientiert ausgebildet werden (vgl. Hesse & Latzko, 2009). Spätestens mit dem Beschluss der Kultusministerkonferenz vom 16.12.2004 zur Einführung von Standards für die Lehrerbildung wird beschrieben, welche Anforderungen an das Handeln der Lehrerinnen und Lehrer gestellt werden. So heißt es im Kompetenzbereich 7 ‚Beurteilen', dass die Lehrkräfte Lernvoraussetzungen und Lernprozesse von Schülerinnen und Schülern diagnostizieren, Schülerinnen und Schüler gezielt fördern und die Lernenden beraten sollen (KMK, 2004). Unter anderem wird für den praktischen Ausbildungsabschnitt in diesem Kompetenzbereich gefordert, dass die Lehrkräfte Entwicklungsstände, Begabungen, Lernpotentiale, Lernhindernisse und Lernfortschritte erkennen und spezielle Fördermöglichkeiten einsetzen sollen (KMK, 2004, 11).

Die geforderte Kompetenzentwicklung bei Lehrpersonen kann allerdings erst erfolgreich sein, wenn vorausgesetzt werden kann, dass Lehrkräfte geeignete Diagnoseinstrumente entweder selber entwickeln können (dies also in ihrer Ausbildung oder in Fortbildungen gelernt haben) oder ihnen wissenschaftlich erprobte Diagnoseinstrumente zur Verfügung gestellt werden und sie damit umgehen können (vgl. Übersicht in Paradies, Linser & Greving, 2009, 63; KMK, 2004). Ein Blick in die deutsche Schullandschaft offenbart, dass Lehrkräfte der Unterrichtsfächer Mathematik und Deutsch relativ breiten Zugang zu solchen Diagnoseinstrumenten besitzen (vgl. Übersicht in Paradies, Linser & Greving, 2009; Kliemann, 2008; Becker, Horstkemper, Risse, Stäudel, Werning & Winter, 2006). Naturwissenschaftliche Fächer, die Physik eingeschlossen, scheinen diesbezüglich offenbar noch nicht so weit entwickelt zu sein. Es gibt kaum Verfahren, mit denen eine Schülerdiagnose auf Klassen- oder auf Individualniveau valide ermöglicht werden kann. Die Diskrepanz zwischen der Forderung, dass Lehrkräfte diagnostische Fähigkeiten ausbilden sollen und der Tatsache, dass die hierzu notwendigen diagnostischen Hilfsmittel für den Physikunterricht kaum existieren, wird in dieser Arbeit zum Anlass genommen, ein für die Schulpraxis geeignetes Instrument für eine Schülerdiagnose im Physikunterricht zu entwickeln.

Die aktuelle Entwicklung vom Halbtagsunterricht in den Ganztagsunterricht (vgl. Holtappels, 2004), stellt einen wichtigen Beitrag für die Zielsetzung dieser Arbeit dar. Denn mehr Lernzeit bietet in den beteiligten Schulen die Möglichkeit, die Schülerinnen und Schüler umfangreicher zu fördern. Dies allerdings setzt ebenfalls eine ausreichende Diagnose mit geeigneten, standardisierten Diagnoseinstrumenten voraus (vgl. Helmke, 2009c), die für den Physikunterricht selten vorzufinden sind (z. B. Teilaufgaben der PISA 2006-Studie).

Eingebettet in das sogenannte Ganz-In-Projekt (finanziert von der Stiftung Mercator), in dem ausgewählte Gymnasien in NRW in ihrem Ausbau zum Ganztagsgymnasium von den Ruhr-Allianz-Universitäten (Ruhr-Universität Bochum, Technische Universität Dortmund (IfS) und die Universität Duisburg-Essen) begleitet werden (Berkemeyer, Bos, Holtappels, Meetz & Rollett, 2010), versucht diese Arbeit den Projektschulen und anderen Gymnasien in NRW ein geeignetes Diagnoseinstrument für den Physikunterricht auf seine Schulpraxistauglichkeit zu prüfen und für die Anwendung im Unterricht anzubieten. Das Instrument soll die Diagnose der Konzeptbildung von Schülerinnen und Schülern im laufenden Unterrichtsprozess ermöglichen. Es muss also mit geringer Vorbereitung einsetzbar sein und zuverlässige Aussagen ermöglichen. Testinstrumente eignen sich nicht, da sie eine zeitintensive Vorbereitung und Auswertung benötigen und deshalb nicht ad hoc einsetzbar sind. Spontane Befragungen der Lernenden zum Erleben ihres Unterrichtserfolgs sind zwar schnell einsetzbar und auszuwerten, aber zu ungenau und kurze einzelne Testaufgaben bezüglich der zu überprüfenden Fähigkeiten zu begrenzt. Beide Verfahren eignen sich außerdem nicht, physikalische Konzepte der Lernenden abzubilden. Wertvolle Informationen über die Konzeptentwicklung der eigenen Schülerinnen und Schüler liefern dagegen Concept Maps (Begriffsnetze). Sie können nicht nur Lernhilfe sein, sondern der Lehrkraft auch für die Schülerdiagnostik, um die es in dieser Arbeit primär geht, brauchbare Hinweise geben. Die Herausforderung im Einsatz von Concept Maps ist es, ein für die Schulpraxis geeignetes Aufgaben- und Bewertungsformat zu entwickeln. Die Praxistauglichkeit des Diagnoseinstruments Concept Map soll in dieser Arbeit über die Praktikabilität ihres Einsatzes und über die Diagnosefähigkeit der Lehrkräfte eingeschätzt werden.

In dieser Arbeit werden zunächst die theoretischen Grundlagen der pädagogischen Diagnostik und der Concept Maps erläutert, um anschließend daraus die Ziele und Forschungsfragen mit den Hypothesen ableiten zu können. Daran schließt sich ein kurzer

Exkurs zu Feldstudien an, um eine Einordnung dieser Arbeit in den Stand der Forschung zu ermöglichen. Mit der Vorstellung der Methoden, des Designs und der entsprechenden Analysemethoden werden die Studien erläutert und die Ergebnisse dargestellt. Den Abschluss bilden die Diskussion und die Zusammenfassung mit einem Ausblick auf weitere Forschung, die sich aus den Ergebnissen entwickeln lässt.

Abbildung 1.1 gibt einen Überblick über die Kernelemente dieser Arbeit.

Abbildung 1.1. Thematische Kernaspekte dieser Arbeit.

2 Theoretischer Hintergrund

In diesem Kapitel wird zunächst der Bereich der *Pädagogischen Diagnostik* im Kontext Schule geklärt. Im Anschluss daran verbindet das Zwischenfazit diesen Abschnitt mit der Funktion von *Concept Maps* in Forschung und Schule.

Die Zusammenfassung dieser beiden übergeordneten Themengebiete beschließt das Kapitel und bildet den Übergang für die daraus abgeleiteten Ziele und Forschungsfragen des folgenden Kapitels.

2.1 Pädagogische Diagnostik

Die für diese Arbeit relevanten Aspekte umfassen die *Begriffsklärungen*, die von Psychologischer Diagnostik bis zu Diagnostik und Diagnose reichen, die Darstellung *Diagnostischer Theorien*, die Beschreibung *Diagnostischer Methoden* und der *Diagnosekompetenz* von Lehrerinnen und Lehrern.

2.1.1 Pädagogische Diagnostik und Diagnose

Begriffe: Psychologische Diagnostik - Pädagogisch-psychologische Diagnostik - Pädagogische Diagnostik

In Anlehnung an Ingenkamp und Lissmann (2008) und Lukesch (1994) lässt sich das Untersuchungsfeld in die Bereiche

- Psychologische Diagnostik (siehe z. B. Jäger & Petermann, 1995; Schmidt-Atzert & Amelang, 2012) und
- Pädagogisch-psychologische Diagnostik und Pädagogische Diagnostik (siehe z. B. Tent & Stelzl, 1993; Langfeldt & Trolldiener, 1993) gliedern.

Der wesentliche Unterschied dieser beiden Bereiche besteht darin, vor welchem Hintergrund diagnostiziert wird und mit welchem Gegenstand sich die jeweiligen Bereiche beschäftigen. In der psychologischen Diagnostik werden Arbeitsfelder angesprochen, die in der pädagogisch-psychologischen und pädagogischen Diagnostik nicht im Fokus stehen. Beispielsweise wird die psychologische Diagnostik im Bereich der klinischen oder neuropsychologischen Diagnostik eingesetzt, bei der es um die Erfassung von Persönlichkeitsmerkmalen, z. B. Beeinträchtigungen der psychischen Persönlichkeit durch eine Erkrankung, geht (Pospeschill & Spinath, 2009, Paradies, Linser & Greving, 2009). In der pädagogisch-psychologischen und pädagogischen Diagnostik werden Lernende

betrachtet und diese beispielsweise hinsichtlich einer Leistungsdiagnostik untersucht. Es geht darum, dass diagnostische Tätigkeiten durchgeführt werden, um Aussagen zu menschlichem Verhalten tätigen zu können. Dies geschieht in Lehr,- Lern- und Erziehungssituationen, wie sie z. B. in Schulen, in beruflicher Aus- und Weiterbildung oder in der Erziehungsberatung zu finden sind. Die Methoden der pädagogisch-psychologischen und pädagogischen Diagnostik stammen aus der psychologischen Diagnostik (Ingenkamp & Lissmann, 2008; Leutner, 2001), sodass in dieser Hinsicht eine Überschneidung zwischen pädagogisch-psychologischer und pädagogischer Diagnostik mit der psychologischen Diagnostik vorhanden ist (Lukesch, 1994). Ingenkamp und Lissmann (2008) betonen allerdings, dass die pädagogisch-psychologische und pädagogische Diagnostik nicht aus der psychologischen Diagnostik entstanden sind und schon immer eigenständig waren. Auch Lukesch (1994) ergänzt, dass die psychologisch-pädagogische und die pädagogische Diagnostik nicht vollständig über die Beschreibung der psychologischen Diagnostik erfasst werden können.

Inhaltliche Merkmale, die die pädagogisch-psychologische von der pädagogischen Diagnostik unterscheiden würde, sind nicht ersichtlich, sodass die sprachliche Trennung nicht begründet werden kann (vgl. u. a. Leutner, 2001). Beide Begriffe werden in der Lehr-Lern-Psychologie synonym verwendet (vgl. u. a. Leutner, 2001; Ingenkamp & Lissmann, 2008).

Die folgende Definition der pädagogischen Diagnostik[1] von Ingenkamp und Lissmann (2008, 13) wird als die für diese Arbeit relevante Definition genutzt:

„[…] Pädagogische Diagnostik umfasst alle diagnostischen Tätigkeiten, durch die bei einzelnen Lernenden und den in einer Gruppe Lernenden Voraussetzungen und Bedingungen planmäßiger Lehr- und Lernprozesse ermittelt, Lernprozesse analysiert und Lernergebnisse festgestellt werden, um individuelles Lernen zu optimieren. […]“.

Begriffe: Diagnostik und Diagnose

Nach Schadé (2002, 1073) wird Diagnose in der Medizin als „Erkennung einer bestimmten Krankheit auf Grund der Beschwerden (Symptome) und Krankheitszeichen nach ärztlicher Untersuchung.“ bezeichnet. Diagnostik sind „Alle auf die Erkennung eines Krankheitsgeschehens als definierte nosologische Einheit gerichteten Maßnahmen.“ Prognose die „Vorhersage einer künftigen Entwicklung auf Grund einer kritischen Analyse

[1]Für diese Arbeit wird vor dem Hintergrund der vorangegangen Erläuterungen durchgängig die Bezeichnung pädagogische Diagnostik genutzt.

des gegenwärtigen Zustandes […].“ (ebenda, 1200). Diagnostik umfasst danach eine Reihe von Maßnahmen, die zur Erstellung einer Diagnose und einer Prognose führen.

Nicht nur Ärztinnen und Ärzte betreiben Diagnostik an Menschen und erstellen eine Diagnose, Pädagoginnen und Pädagogen und Psychologinnen und Psychologen beschäftigen sich ebenfalls in ihren Arbeitsfeldern mit der Diagnostik von Personeneigenschaften und -merkmalen. Psychologinnen und Psychologen betreiben u. a. als ärztliches Fachpersonal klinische Diagnostik, um Krankheiten festzustellen. Pädagogisches Fachpersonal und Lehrerinnen und Lehrer beschäftigen sich mit der Diagnostik im Kontext pädagogischer Fragestellungen, die Schülerinnen und Schüler im Kindergarten, Grundschul-, Jugend- oder im Erwachsenenalter (z. B. Pädagogische Fachkräfte und Lehrpersonen in der Berufs- und Weiterbildung) betreffen. Nach Jäger und Petermann (1995) besteht die Diagnostik, bezogen auf den pädagogischen Bereich, „im systematischen Sammeln und Aufbereiten von Informationen mit dem Ziel, Entscheidungen und daraus resultierende Handlungen zu begründen, zu kontrollieren und zu optimieren. […]“ (Jäger & Petermann, 1995, 11). Dies führt dazu, dass damit pädagogisch-psychologische Charakteristika von Merkmalsträgern erkannt und die in der Diagnostik gewonnenen Daten zu einem Urteil (Diagnose, Prognose) integriert werden können (Jäger & Petermann, 1995).

Diagnostische Theorien und Diagnostische Methoden in der pädagogischen Diagnostik

Ein diagnostisches Vorgehen in der pädagogischen Diagnostik wird durch die Zielsetzung und den Zweck der geplanten Diagnose festgelegt (Ingenkamp & Lissmann, 2008). Je nach Grund des Diagnoseprozesses, muss eine geeignete *Theorie* ausgewählt werden. Die Klassifizierung der Diagnosetheorie lässt sich nach Siemes (2008) in Status- und Prozessdiagnostik unterscheiden (vgl. u. a. Leutner, 2001; Ingenkamp & Lissmann, 2008). Durch weitere diagnostische Theorien, die an dieser Stelle nicht weiter diskutiert werden, können die Status- und die Prozessdiagnostik weiter charakterisiert werden (vgl. Abb. 2.1).

Bei der *Statusdiagnostik* steht das Erfassen des Zustandes einer Person mit dem Ziel im Vordergrund, in einer ganz bestimmten Situation zu selegieren (z. B. Gutachten für Schullaufbahnberatungen im Sinne einer Leistungsdiagnostik). Ein Test zur Erfassung des Fachwissens in einem speziellen physikalischen Inhaltsbereich eignet sich beispielsweise zur Statusdiagnostik. Im Rahmen dieser Diagnostiktheorie wird auf die bei den

Schülerinnen und Schülern gefundenen Defizite fokussiert. Die Statusdiagnostik wird auch Selektionsdiagnostik genannt. In der *Prozessdiagnostik* werden die Prozesse und Aspekte untersucht, mit deren Kenntnis eine Veränderung des Verhaltens und des Erlebens einer Person eingeleitet werden kann (Siemes, 2008). Es geht darum, den bereits vorhandenen Kenntnisstand einer Person zu erfassen. Das Wissen über Verhaltensabläufe (Prozesse) ermöglicht es, geeignete Maßnahmen einzuleiten und auf die Person auszurichten. Bei einem länger dauernden Diagnostikprozess müssen diese Prozesse regelmäßig überprüft werden, um die eingeleiteten Maßnahmen für die betreffende Person anzupassen. Demzufolge ist die Prozessdiagnostik gleichzeitig eine Modifikationsdiagnostik. Dieses Vorgehen kann mit der Diagnostik verglichen werden, die eine Lehrperson im laufenden Unterrichtsprozess einsetzen muss, um für die einzelnen Schülerinnen und Schüler einer Lerngruppe den Lernprozess zu modifizieren und ihn den Bedarfen anzupassen.

Beide vorgestellten diagnostischen Theorien können genutzt werden, um persönliche Voraussetzungen und Potenziale der Schülerinnen und Schüler festzustellen und um daraus den eigenen Unterricht angemessen zu adaptieren und individuelle Lernhilfen vorzuschlagen.

Abbildung 2.1 veranschaulicht die Differenzierung nach Siemes (2008), wobei zusätzlich als übergeordnete Kategorie die systemische Diagnostik eingeführt wird, bei der das soziale Gefüge, in dem sich die Personen befinden, untersucht wird. Dieser Diagnostikteil wird in dieser Arbeit nicht weiter diskutiert.

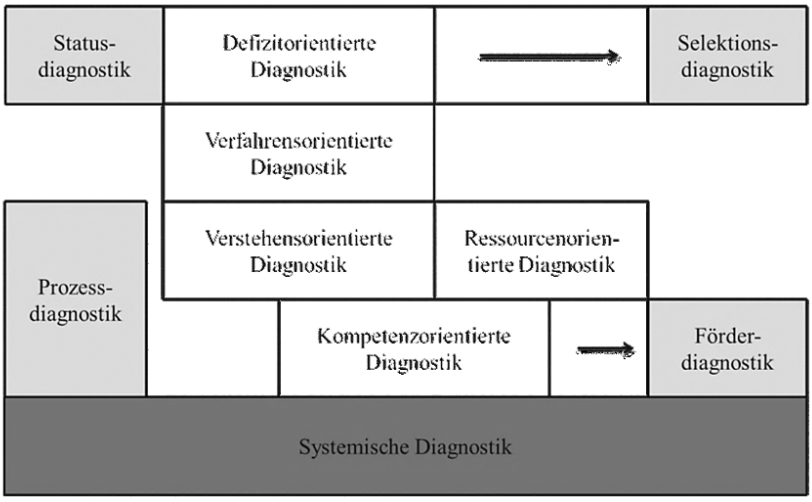


Abbildung 2.1. Zusammenhang der verschiedenen Diagnosetheorien (nach Siemes, 2008, 17).

Entscheidet sich der Diagnostiker für eine der Theorien, erfolgt anschließend die Wahl einer geeigneten *Methode*, um die Diagnose durchzuführen. Die Auswahl der geeigneten diagnostischen Methode charakterisiert ebenfalls den Diagnostikprozess (vgl. Wild & Krapp, 2006).

Die im Folgenden vorgestellten Methoden können generell für die Erfassung von individuellen Personenmerkmalen (wie z. B. Vorwissen von Schülerinnen und Schülern zu einem bestimmten Bereich) und Umweltmerkmalen hinsichtlich der Lehr-Lern-Umwelt der Personen (beispielsweise inwiefern eine familiäre Unterstützung gegeben ist) angewendet werden. Je nach Autor können beispielsweise Methoden der Beobachtung sowohl für die Personendiagnostik als auch für die Umweltdiagnostik eingesetzt werden (vgl. Wild & Krapp, 2006; vgl. Ingenkamp & Lissmann, 2008).

Ingenkamp und Lissmann (2008) schlagen als grobe Orientierung Methoden der *Verhaltensbeobachtung, Befragung* und *Testung* vor (vgl. auch Schmidt-Atzert & Amelang, 2012). Während Verhaltensbeobachtungen durch schriftliche Fremd- oder Selbstbeurteilungsbögen oder über das Medium Video erfasst werden können, kann über Befragungsmethoden, wie das Interview oder in Gesprächssituationen, nach Einflüssen auf das Lernverhalten gefragt werden (Lukesch, 1994). Testverfahren, wie der Einsatz eines schriftlichen Multiple-Choice-Tests, ermitteln z. B. die Rechenfähigkeit im Bereich Bruchrechnen einer Schülerin oder eines Schülers. Schülerleistungen können über die klassischen Wege der mündlichen und schriftlichen Prüfungen und Schulleistungstests beurteilt werden (Pospeschill & Spinath, 2009).

Die genaue Darbietungsform der diagnostischen Methode (also einer Verhaltensbeobachtung, Befragung oder einer Testung) hängt von der Zielgruppe und dem Ziel der Diagnose ab, sodass die Formen auf diese Aspekte angepasst werden müssen. Eignet sich ein Lückentext für die Erprobungsstufe der Klasse 6 zur Erfassung von Sprachkenntnissen, kann durch einen Diagnosebogen in Form einer Checkliste das soziale Arbeitsverhalten einzelner Schülerinnen und Schüler in diesem Jahrgang ermittelt werden. Weitere Darbietungsformen sind u. a. (vgl. u. a. Ingenkamp & Lissmann, 2008; Pospeschill & Spinath, 2009)

- Checklisten
- Beobachtungsbögen
- Videoanalysen
- Interviews
- Fragebögen
- Multiple-Choice-Tests
- Kurzantworten
- Lückentexte
- Zuordnungen
- Essays

- Portfolios
- Diagnosebögen
- Mind Maps
- Concept Maps

Die Auflistung möglicher Darbietungsformen der diagnostischen Methoden kann in verschiedenen diagnostischen Theorien wie der Status- oder Prozessdiagnostik eingesetzt werden. Eine klare Zuordnung, welche Darbietungsform genau zu welcher Methode und zu welcher Diagnosetheorie gehört, gibt es nicht. Viele Formen sind in verschiedenen Schulfächern einsetzbar. Im Bereich Unterrichtsmaterialien/Diagnosehilfsmittel gibt es für deutschsprachigen Unterricht wissenschaftlich abgesicherte Diagnoseverfahren/-instrumente bislang nur für die Fächer Mathematik, Deutsch und Englisch (vgl. Übersicht in Paradies, Linser & Greving, 2009). Die Hamburger Schreib-Probe (May, 2007) für das Fach Deutsch oder der Rechentest +9 (Bremm & Kühn, 1992) für Mathematik sind diagnostische Tests. Für das Fach Englisch werden von den Schulbuchverlagen wie Cornelsen oder Diesterweg Tests angeboten. Physiklehrerinnen und Physiklehrer können bislang auf keine zuverlässigen Diagnoseverfahren zurückgreifen, die individuelle Aussagen über die Schülerinnen und Schüler machen können. Vergleichsstudien wie PISA 2006 mit dem Schwerpunkt Naturwissenschaften oder die Evaluation der Bildungsstandards im EsNaS-Projekt (Walpuski, Kauertz, Kampa, Fischer, Mayer, Sumfleth & Wellnitz, 2010) können keine diagnostische Rückmeldung auf Individualbasis leisten. Physiklehrkräfte müssen derzeit Diagnoseinstrumente nach eigenen Kriterien entwickeln.

2.1.2 Diagnosekompetenz von Lehrkräften

Nach der Erläuterung des Bereichs der pädagogischen Diagnostik, schließt sich eine detaillierte Beschreibung der Diagnosekompetenz von Lehrkräften an.

Diagnostische Aufgaben von Lehrkräften

Neben dem Lehren von Fachinhalten sollen Lehrerinnen und Lehrer nach KMK (2004) ebenfalls Erziehen, Innovieren und Beurteilen. Beurteilen bedeutet, die Fähigkeiten von Schülerinnen und Schülern zutreffend einzuschätzen. Lehrerinnen und Lehrer müssen diagnostisch tätig werden. Jäger (2009) nennt hierzu verschiedene diagnostische Aufgaben, die in unterschiedlichen Unterrichtssituationen auftreten können. Es müssen Zensuren vergeben werden, der eigene Unterricht muss bewertet werden, es müssen Aussagen über den Grad getroffen werden, wie Lernziele erreicht wurden, den Schülerinnen und Schülern

Rat bei der Fächerwahl gegeben werden oder es müssen Aussagen über das Klassenklima getroffen werden (Jäger, 2009). Es wird deutlich, dass von Lehrkräften diagnostische Aufgaben auf verschiedenen Ebenen bewältigt werden müssen. Langfeldt (2006) (nach Hesse & Latzko, 2009) schlägt eine Kategorisierung der diagnostischen Aufgaben in drei Ebenen vor:

a) auf individueller Ebene (ein Physiklehrer will z. B. die Defizite eines Schülers beurteilen),

b) auf Klassenebene (die Lehrkraft stellt die Unterschiede zwischen den Schülerinnen und Schülern fest) und

c) auf institutioneller Ebene (die Lehrkraft schreibt z. B. Zeugnisse).

Lehrkräfte führen diese Diagnostikprozesse mehrheitlich intuitiv durch. Die schulische Diagnostik ist häufig unsystematisch und sie bewegt sich auf einer Ebene der informellen subjektiven Einschätzung (Schrader, 2001). Wenn Lehrkräfte diagnostische Aufgaben erfolgreich lösen sollen, setzt dies diagnostische Kompetenz voraus.

Der Begriff der Diagnostischen Kompetenz und seine Komponenten

Die Fähigkeit, Merkmale von Schülerinnen und Schülern zutreffend einzuschätzen und die diagnostischen Aufgaben in der Schule bzw. im pädagogischen Umfeld adäquat auszuüben, kann allgemein als diagnostische Kompetenz oder Diagnosekompetenz von Lehrkräften bezeichnet werden (vgl. u. a. Artelt & Gräsel, 2009; Schrader, 2001; Gläser-Zikuda, 2010; Anders, Kunter, Brunner, Krauss & Baumert, 2010). Die Diagnosekompetenz wird als eine der vier Schlüsselkompetenzen von Lehrkräften benannt (Weinert, 1998 in Anders et al., 2010 und vgl. Weinert, 2000). Spätestens seit der PISA 2000-Studie wurden der Ruf und die Diskussionen um den Begriff Diagnosekompetenz immer stärker, der trotz vermehrter Forschungsbemühungen bis heute vage geblieben ist (vgl. Hesse & Latzko, 2009; Helmke, 2009a). Der Begriff ist vage, weil eine präzise Operationalisierung, die mehr als eine Komponente der Diagnosekompetenz messbar macht, bis heute in der Forschung und in den praxisnahen Studienseminaren und Schulen kaum gelungen ist (siehe Abschnitt 2.1.2 Einordnung der diagnostischen Kompetenz in die aktuelle Forschungslage).

Die momentan gängigste Definition der Diagnosekompetenz für den deutschsprachigen Raum stammt aus dem Bereich der Lehr-Lern-Psychologie (vgl. Schrader & Helmke, 1987). Bereits Ende der 80er Jahre wurde mit Schraders Arbeit „Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und

Effektivität des Unterrichts“ (Schrader, 1989) angedeutet, wie die Diagnosekompetenz von Lehrkräften messbar gemacht werden kann. Der Blick in die Forschungsjahre danach verdeutlicht, dass der Versuch, Komponenten des Konstrukts ‚Diagnosekompetenz‘ näher zu umschreiben, bis heute noch nicht zum Ziel geführt hat. Das Konstrukt ‚Diagnosekompetenz‘ wird hauptsächlich über Handlungen und Fallbeispiele beschrieben, die die Lehrkräfte ausüben oder bewerten sollen (vgl. Studien von Cappell & von Aufschnaiter, 2011 & 2012; Haschke-Hirth & Kuhle, 2010; Komorek & Michaelis, 2011). Diese Arbeiten gehen allerdings wenig auf die Thematik der Messung von Diagnosekompetenz ein. Eine klare Definition, wie sie bei der Beschreibung von Schülerkompetenzen vorgenommen wurde (z.B. in PISA als Problemlösekompetenz oder Sprachkompetenz operationalisiert, vgl. Baumert, Klieme, Neubrand, Prenzel, Schiefele, Schneider, Stanat, Tillmann & Weiß, 2001), kann für das Konstrukt ‚Diagnosekompetenz‘ nicht festgestellt werden. Deshalb lehnt sich diese Arbeit an die Untersuchungen von Schrader und Helmke (1987) an, in denen eine Operationalisierung und konkrete Messmethoden für die Erfassung der diagnostischen Kompetenz von Lehrkräften vorgeschlagen werden. Ihre Arbeiten gelten als grundlegend und werden in vielen Studien genutzt (siehe auch Abschnitt 2.1.2 Einordnung der diagnostischen Kompetenz in die aktuelle Forschungslage).

Die diagnostische Kompetenz wird nach aktueller Forschungslage momentan über die sogenannte Diagnosegenauigkeit bestimmt. Hierbei kann nach Schrader und Helmke (1987) davon ausgegangen werden, dass die Diagnosegenauigkeit die Diagnosekompetenz einer Person widerspiegelt. Nach Helmke, Hosenfeld und Schrader (2004) erfordert angemessenes und effektives Unterrichten eine Abstimmung der Schülerfähigkeiten mit den von der Lehrkraft angebotenen Unterrichtsarrangements. Dies bedeutet, dass eine realistische Einschätzung der Fähigkeiten benötigt wird, die von der Diagnose abhängt (auch als Diagnosegenauigkeit bekannt). Die Diagnosegenauigkeit lässt sich (übergeordnet als Diagnosekompetenz bezeichnet) über die Faktoren der *Rangordnungskomponente, Niveaukomponente* und *Streuungskomponente* messen (vgl. Helmke, 2009a; Helmke, Hosenfeld & Schrader, 2004; Schrader, 1989).

Die Rangordnungskomponente zeigt, inwiefern eine Lehrkraft in der Lage ist, ihre Schülerinnen und Schüler erfolgreich in eine Rangfolge zu bringen. Dabei wird die Testleistung der Schülerinnen und Schüler zu einem bestimmten Themengebiet in eine Rangfolge gebracht und mit der Rangfolge korreliert, die die Lehrkraft über dieselbe Lerngruppe eingeschätzt hat. Je höher die Korrelation ist, desto ähnlicher sind beide

Rangfolgen und entsprechend genauer die Einschätzung der Lehrkraft. Dies wird mit dem Begriff der personenbezogenen Rangordnungskomponente beschrieben (Schrader, 1989). Die Studien von Helmke, Hosenfeld und Schrader (2004) und Schrader (1989) haben gezeigt, dass ebenfalls eine aufgabenbezogene Rangordnungskomponente messbar ist, bei der das Bilden der Rangordnung erhalten bleibt. In diesem Fall schätzen Lehrkräfte die Schwierigkeiten von Aufgaben, bilden darüber eine Rangfolge und diese wird mit der empirischen Aufgabenschwierigkeit, die als Rangfolge vorliegt, verglichen.

Die Niveaukomponente lässt sich ebenfalls in eine personenbezogene und eine aufgabenbezogene Niveaukomponente gliedern. Durch sie wird ausgesagt, ob die Personen oder die Aufgaben, die von der Lehrkraft eingeschätzt werden sollen, im Mittel über- oder unterschätzt wurden. Bei Schrader (1989, 87) heißt es: „[…] Er [*gemeint ist der berechnete Wert der personenbezogenen Niveaukomponente*] gibt die Differenz zwischen der mittleren Einschätzung eines Lehrers und der mittleren Leistung der von ihm eingeschätzten Schüler an. […]“. Wenn dieses gerichtete Maß einen Wert größer Null annimmt, hat eine Lehrperson das Leistungsniveau überschätzt. Ein Wert kleiner Null offenbart eine Unterschätzung und ein Wert gleich Null kennzeichnet eine exakte Einschätzung der Lehrkraft. Analog verhält es sich bei der aufgabenbezogenen Niveaukomponente.

Die letzte Komponente der Diagnosegenauigkeit ist die sogenannte (personenbezogene oder aufgabenbezogene) Streuungskomponente. Durch sie kann, ähnlich wie bei der Niveaukomponente, ein Wert berechnet werden, der im Fall der personenbezogenen Streuungskomponente das „[…] Verhältnis zwischen der Streuung der Einschätzung eines Lehrers und der dazu korrespondierenden Leistungsstreuung bei den Schülern [*angibt*…]. Werte größer als 1 kennzeichnen eine Überschätzung, Werte kleiner als 1 eine Unterschätzung und ein Wert von 1 eine exakte Einschätzung der Leistungsstreuung […]“ (Schrader, 1989, 87f.). Für die exakte Berechnung der Rangordnungs-, Niveau- und Streuungskomponente wird an dieser Stelle auf Schrader (1989) und Schrader und Helmke (1987) verwiesen.

Nach Abs (2007) ist diese Konzeption der Diagnosegenauigkeit nicht vollends zufriedenstellend, da die didaktische Relevanz unklar bleibt. Ebenfalls muss berücksichtigt werden, welche Bedeutung die Genauigkeit einer Lehrerdiagnose im pädagogischen Alltag hat. Tatsächlich muss eine diagnostische Kompetenz von Lehrkräften vorhanden sein, um den Unterricht auf die Schülerinnen und Schüler abzustimmen und um individuell fördern zu können (vgl. Helmke, 2009b). Die Diagnosen müssen im Unterrichtsverlauf aber nicht

immer genau sein, wenn eine Lehrkraft sich „[…] der Ungenauigkeit, Vorläufigkeit und Revisionsbedürftigkeit seiner Urteile bewusst ist. […]“ (Weinert & Schrader, 1986, 18).

Einordung der diagnostischen Kompetenz in die aktuelle Forschungslage

An dieser Stelle wird ein Überblick über den Forschungsstand zur Thematik der Diagnostischen Kompetenz von Lehrkräften gegeben. Es wird verdeutlicht, warum die beschriebene Modellierung und Operationalisierung der Diagnosekompetenz in Form der Diagnosegenauigkeit und ihrer Komponenten nach Schrader und Helmke (1987) als Forschungsgrundlage für diese Arbeit gewählt wird.

Bei der Zuordnung der bereits existenten Untersuchungen zur diagnostischen Kompetenz von Lehrkräften zeichnet sich ab, dass dieses Gebiet vorrangig von Lehr-Lern- und Sozial-Psychologen untersucht wird. Es können Untersuchungen aus dem englischsprachigen Bereich zur sogenannten Judgement-Accuracy (Hoge, 1983; Hoge & Coladarci, 1989) und im deutschsprachigen Raum zur diagnostischen Kompetenz (Schrader, 1989; Südkamp, Möller & Pohlmann, 2008) herangezogen werden. Schrader und Helmke (1987) unterscheiden zunächst zwei Gruppen empirischer Arbeiten zur Diagnosekompetenz:

> a) Studien, die unter deskriptiver Zielsetzung versuchen, „[…] Aussagen über die Vorhersagekraft von Lehrerurteilen für verschiedene Aspekte der Leistungsfähigkeit von Schülern zu gewinnen (Brennan & Redding, 1985; Hopkins, George & Williams, 1985 […]) und mit anderen Prädiktoren zu vergleichen“ (Schrader & Helmke, 1987, 29) und
>
> b) Studien, die Faktoren identifizieren, die Lehrkräfte in ihren Urteilen und in ihrer Urteilsgenauigkeit beeinflussen.

Diese Einteilung kann durch Untersuchungen der jüngeren Zeit ergänzt werden. Es ist unter anderem den Ergebnissen der PISA 2000-Studie geschuldet, dass das Interesse an der Untersuchung der Diagnosekompetenz von Lehrkräften in den letzten Jahren im deutschsprachigen Raum gestiegen ist. PISA 2000 offenbarte an einer kleinen Stichprobe eine optimierbare Diagnosekompetenz von Lehrkräften (vgl. u. a. Helmke, 2009a). Daraufhin sind verschiedene Untersuchungen im deutschsprachigen Raum durchgeführt worden. Die Tendenz in der Forschung geht dahin, das Konstrukt der Diagnosekompetenz zu operationalisieren und neben der Modellierung nach Schrader und Helmke (1987) weiter zu präzisieren. Durch Forschungsprojekte wie UDiKom in der Psychologen-Gruppe um Leutner und Wirth (vgl. Haschke-Hirth & Kuhle, 2010), das Projekt „LUV – Lernen

aus Unterrichtsvideos“ (Seidel & Prenzel, 2007) oder das Lehrerausbildungsprojekt „OLAW zur Entwicklung von Diagnose- und Förderkompetenz“ (Komorek & Michaelis, 2011) wurden Maßnahmen und Instrumente entwickelt, die die Diagnosekompetenz von Lehrkräften entwickeln sollen. Diagnosekompetenz wird vorrangig durch die konkreten Handlungen beschrieben, die die Lehrkraft durchführt. Die Messung der Diagnosekompetenz erfolgt jedoch in diesen Arbeiten nicht nach der Definition von Schrader und Helmke (1987).

Schrader und Helmke (1987) untersuchen die Diagnosekompetenz von Mathematiklehrkräften. Danach überschätzen Lehrkräfte z. B. die Diagnosegenauigkeitskomponente Leistungsstreuung in ihren Klassen mehrheitlich. Die Ergebnisse zeigen zusätzlich, dass der „[…] leistungssteigernde Effekt von Strukturierungshilfen von der diagnostischen Kompetenz […] abhängt“ (Helmke, 2009a, 132). Ergänzt mit zusätzlichen Strukturierungshilfen korreliert hohe diagnostische Kompetenz mit Lernerfolg (Helmke, 2009a). In der VERA-Studie schätzt die Mehrheit der Grundschullehrkräfte für Mathematik die Schwierigkeit von Aufgaben angemessen ein (Helmke, Hosenfeld & Schrader, 2004). Allerdings erraten 10% der Lehrkräfte die Aufgabenschwierigkeit mit wenig Erfolg. Jüngere Studien zeigen, dass die Diagnosekompetenz bei Lehrerinnen und Lehrern verschiedener Fächer (Deutsch, Mathematik und Englisch, vgl. McElvany, Schroeder, Hachfeld, Baumert, Richter, Schnotz, Horz & Ulrich, 2009) und Schulformen (Karing, 2009) schwach bis moderat ausgeprägt war. In allen genannten Studien wurde die Diagnosegenauigkeit über die Operationalisierung nach Schrader und Helmke (1987) gemessen.

Diese Befunde decken sich mit der Metaanalyse zur Diagnosegenauigkeit von Hoge und Coladarci (1989). Bei einer akzeptablen mittleren Diagnosegenauigkeit war die Varianz groß. Es wurden die Korrelationen zwischen den Lehrerurteilen und der mit einem Test erbrachten Schülerleistung ermittelt. Die Werte streuten individuell zwischen 0.28 und 0.92 und lagen im Median bei $r = 0.66$.

Abs (2007) stellt in seiner Arbeit mit dem Titel „Überlegungen zu einem Kompetenzmodell für die Erfassung der Diagnosekompetenz bei Lehrerinnen und Lehrern“ an und versucht dabei zunächst über konkrete Anforderungssituationen im Rahmen der diagnostischen Aufgaben einen Zugang zur Thematik zu erhalten. Jedoch ist ihm bislang noch kein Modell zur Beschreibung der Diagnosekompetenz gelungen, das die diagnostische Kompetenz einer Lehrperson in Kompetenzstufen, ähnlich der Modellierung in PISA, durch empirische Untersuchungen zeigen kann. Karst (2012) identifiziert

Elemente eines Kompetenzmodells zu diagnostischen Urteilen von Grundschullehrkäften und orientiert sich ebenfalls an Schrader und Helmke (1987). Die Modellierung der Diagnosekompetenz von Physiklehramtsstudierenden von Rath und Reinhold (2014) ist momentan noch in einem Entwicklungsprozess.

Für die Bestimmung der Genauigkeit der Diagnosen werden neben der Definition der drei Komponenten (Rangordnungs-, Niveau- und Streuungskomponente) nach Schrader und Helmke (1987) keine Alternativen genannt.

2.1.3 Zwischenfazit

Diagnostische Tätigkeiten gehören zu den Schlüsselaufgaben einer Lehrperson. Ohne eine Diagnostik der Schülerinnen und Schüler können beispielsweise adaptive Maßnahmen im Unterricht nur intuitiv und nicht valide und reliabel vorgenommen werden. Im Sinne einer pädagogischen Diagnostik erkennen, beurteilen, bewerten Lehrpersonen und sie geben Empfehlungen in verschiedenen Situationen und auf verschiedenen Ebenen. Es kann *selektiv* diagnostiziert werden, um individuelle Schülerleistungen zu erfassen und prozessbegleitend (somit *modifizierend*), um die eingeleiteten individuellen Maßnahmen und den Unterricht auf die Schülerbedürfnisse zu adaptieren. Der Lehrkraft stehen verschiedene *diagnostische Methoden* zur Verfügung, um diese Urteile bilden zu können (z. B. Methoden der Verhaltensbeobachtung, Befragung oder Testung).

Eine erfolgreiche Bewältigung der vielfältigen diagnostischen Aufgaben setzt *diagnostische Kompetenz* voraus.

Unterschiedliche empirische Untersuchungen haben gezeigt, dass Lehrpersonen im deutschsprachigen Raum gering ausgeprägte diagnostische Fähigkeiten aufweisen. Die Messung von Diagnosekompetenz findet in den Untersuchungen hauptsächlich über *Diagnosegenauigkeit* statt, mit der Diagnosekompetenz in den meisten Studien operationalisiert wird. Die Messkomponenten der Diagnosegenauigkeit sind die Niveaukomponente, die Streuungskomponente und die Rangordnungskomponente als Maß für die Genauigkeit eines Lehrerurteils. Die Forschung der letzten Jahre bemüht sich um eine alternative Modellierung und Operationalisierung des Begriffs Diagnosekompetenz. Allerdings ist das momentan gängigste Maß das der Diagnosegenauigkeit. In dieser Studie wird die Diagnosegenauigkeit durch die Korrelation der Rangordnung von Lehrerurteil über die Schülerinnen und Schüler und Testergebnissen in einem Inhaltsbereich bestimmt (siehe Abbildung 2.2).

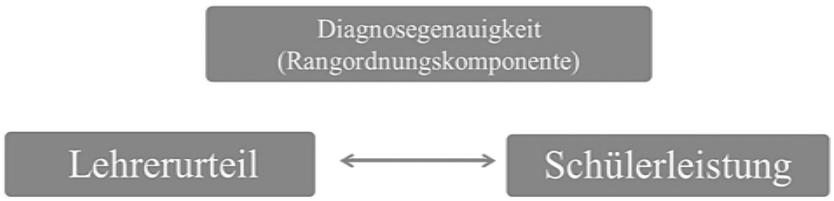


Abbildung 2.2. Zusammenhang der Rangordnungskomponente.

Um pädagogische Diagnostik betreiben zu können, benötigen Lehrkräfte Diagnoseinstrumente. Problematisch erscheint, dass speziell in der Physikdidaktik kaum wissenschaftlich erprobte Instrumente existieren. Lehrkräfte der Physik sind gegenüber Lehrerinnen und Lehrern der Fächer Deutsch und Mathematik in einer benachteiligten Situation. Es ist wünschenswert, diesen Zustand anzugleichen und ein Diagnoseinstrument zu entwickeln, das eine Diagnostik von Schülerlernzuständen im Sinne einer *prozessorientierten Diagnostik* ermöglicht.

2.2 Concept Maps

Dieses Kapitel soll mit Concept Maps ein Diagnoseinstrument vorstellen, mit dem eine effektive Status- und Prozessdiagnostik auf Schülerebene ermöglicht werden kann.

Für das bessere Verständnis wird zunächst erläutert, was Concept Maps sind, um anschließend eine Klassifizierung der Einsatzmöglichkeiten von Concept Maps im pädagogischen Kontext zu präsentieren. Im weiteren Verlauf werden speziell ausgewählte Forschungsergebnisse zum Einsatz mit und zur Qualität über Concept Mapping vorgestellt. Der Einsatz von Concept Maps als Diagnoseinstrument für den Physikunterricht soll dieses Kapitel schließen.

2.2.1 Concept Mapping

Concept Maps sind im deutschsprachigen Raum als Begriffsnetz oder Begriffslandkarte (vgl. u. a. Peuckert, 1999) bekannt. Sie können als eine Möglichkeit angesehen werden, Wissensstrukturen einer Person zu repräsentieren. Ursprünglich wurden Concept Maps von der amerikanischen Wissenschaftlergruppe um Joseph Novak als

Auswertungsverfahren für Interviews eingesetzt, die mit Schülerinnen und Schülern gemacht wurden. Anschließend wurde die Idee dieser Maps theoretisch fundiert und als Diagnoseinstrument von Wissensstrukturen Lernender genutzt (vgl. Novak & Gowin, 1984; Novak, 1990). Parallel zur angloamerikanischen Entwicklung wurde mit der sogenannten Heidelberger-Struktur-Lege-Technik (Scheele & Groeben, 1984) ein ähnliches Verfahren für den deutschsprachigen Raum entwickelt.

Die einfachste Form einer Concept Map sieht vor, dass verschiedene Begriffe (engl. Concepts) eines bestimmten Themengebietes in einer gewissen Form (z. B. hierarchisch oder netzartig) angeordnet und diese Begriffe miteinander über beschriftete Pfeile verbunden werden. Die Beschriftungen der Pfeile, Relation genannt, geben an, welcher Sinnzusammenhang zwischen den Begriffen besteht. Ein Pfeil kann immer nur zwei Begriffe verbinden und den Zusammenhang zwischen diesen beiden Begriffen angeben. Dieses Element ‚Begriff-beschrifteter Pfeil-Begriff' wird als Proposition bezeichnet (vgl. u. a. Behrendt & Reiska, 2001; Haugwitz, 2009; Stracke, 2004). Concept Maps können hierarchisch strukturiert sein, eine Form der Concept Maps, von der in den 80-er Jahren überwiegend ausgegangen wurde (Fischler & Peuckert, 2000). Zu Beginn der Forschung in den 80-er Jahren wurden zunächst Concept Maps aus dem Bereich der Biologie betrachtet, die hierarchisch orientiert waren. Die Ergebnisse „[…] führten bei Novak und Mitarbeitern durchgängig zur Vorstellung, Concept Maps, die die Wissensstruktur von Schülern beschrieben, seien grundsätzlich hierarchisch strukturiert. […]" (Fischler & Peuckert, 2000, 5). Es lässt sich allerdings nach heutigem Forschungsstand nicht mehr begründet erklären, warum einzig die hierarchischen Concept Maps die Wissensstruktur zu einem bestimmten Fachthema angemessen darstellen sollen (Fischler & Peuckert, 2000). Eine Vielzahl von Untersuchungen hat gezeigt, dass Lernende neben hierarchischen auch Concept Maps erstellen, die kettenartig, kreisartig oder netzartig angeordnet sind (vgl. Ruiz-Primo & Shavelson, 1996; u. a. Studien von McClure, Sonak & Suen, 1999; Hucke & Fischer, 2000).

Abbildung 2.3 zeigt ein Beispiel für eine hierarchische Concept Map zum Thema Magnetismus. In Kapitel 2.2.3 wird dargestellt, wie Concept Maps erstellt werden können.

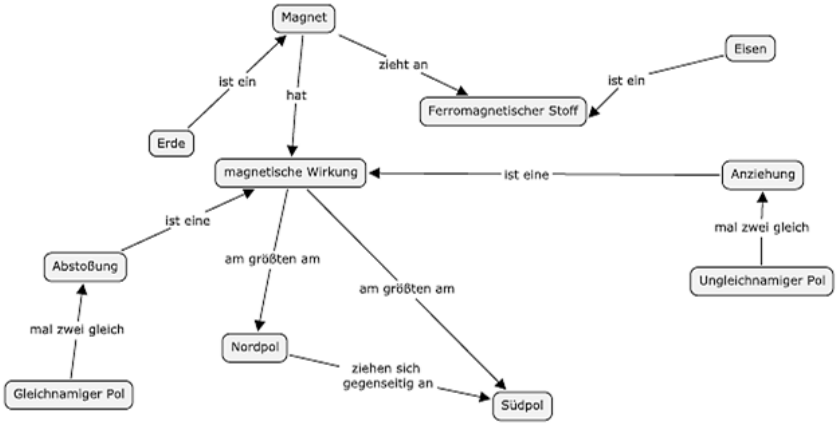


Abbildung 2.3. Beispiel einer Concept Map zum Thema Magnetismus.

2.2.2 Anwendungsmöglichkeiten des Concept Mapping

Ein wesentlicher Vorteil von Concept Maps ist ihr Potenzial zur Repräsentation von Sachstrukturen und den korrespondierenden Wissensstrukturen und Konzeptvorstellungen eines Lernenden. Von der Lehr-Lern-Psychologie werden Concept Maps mehrheitlich als *Lehr- und Lernstrategien* oder in *Kooperationsprozessen* beim gemeinsamen Lernen eingesetzt (vgl. Übersichten in Mandl & Fischer, 2000; Nesbit & Adesope, 2006; u. a. Studien von Renkl & Nückles, 2006; Tergan, 2006).

Stracke (2004) ergänzt die Einsatzoptionen um den Bereich der *Curriculumentwicklung* und *Unterrichtsplanung*. Concept Maps erlauben die Curriculuminhalte zu strukturieren (vgl. Studie von Starr & Krajik, 1990) und geben so einer Lehrperson die Möglichkeit, den eigenen Unterricht zu gliedern. Im Sinne eines Advance Organizers (=Übersicht der wichtigsten Begriffe beispielsweise eines Sachtextes; für weitere Erläuterungen vgl. Ausubel, 1960), der den Lernenden von einer Lehrperson angeboten wird, wird die Struktur des Unterrichts für die Schülerinnen und Schüler transparent gemacht. In dieser Hinsicht ist das Concept Mapping als Lehrmittel bzw. Lehrstrategie einzustufen. Lernende selber können Concept Mapping im Sinne einer Lernstrategie nutzen, wenn das eigene Wissen zu einem bestimmten Themenfeld organisiert und konstruiert werden soll. In Gruppen- oder Partnerarbeit kann z. B. anschließend kooperativ weiter daran gearbeitet werden.

Das für diese Arbeit primär interessierende Anwendungsgebiet ist das Concept Mapping als *Diagnoseinstrument*. Bei der Nutzung von Concept Maps als Diagnoseinstrument kann die Lehrperson Schülervorstellungen und Konzepte der Schülerinnen und Schüler zu einem Thema diagnostizieren und den Zeitpunkt des

Einsatzes im jeweiligen Unterricht bestimmen (vgl. u. a. Jüngst & Strittmatter, 1995). Nach der Unterweisung der Schülerinnen und Schüler in das Erstellen von Concept Maps (über die Wichtigkeit des Einübens siehe auch Sumfleth, Neuroth & Leutner, 2010; Jüngst & Strittmatter, 1995), können Concept Maps zur Vorwissensabfrage und Schülervorstellungserfassung, zur Zwischendiagnose oder am Ende einer Unterrichtseinheit zur Leistungsmessung genutzt werden (vgl. Stracke, 2004). Ein Vergleich von Concept Mapping mit anderen Diagnoseverfahren zeigt, dass der wesentliche Vorteil im geringen Vorbereitungsaufwand und dem diagnostischen Ertrag zu sehen ist. Lehrkräften wird eine schnelle Diagnose der Schülerinnen und Schüler individuell und auf Klassenebene ermöglicht. Will eine Lehrkraft unterrichtsbegleitend diagnostizieren, muss sie im Idealfall die Aufgabenstellung nur einmal entwerfen (z. B. welche Worte in der Concept Map genutzt werden sollen) und sie erhält durch die generierten Maps Material für eine Diagnose der Schülerleistungen. Testverfahren benötigen mehr Vorbereitungsaufwand und das diagnostische Potenzial von Test-Aufgaben kann sich verringern, wenn die Aufgaben mehrmals eingesetzt werden (vgl. u. a. Kauertz & Fischer, 2010). In Abgrenzung zur üblicherweise im Unterricht eingesetzten intuitiven Diagnose (vgl. Schrader, 2001), lässt sich klar herausstellen, dass Concept Maps wie Tests die Diagnose strukturierter herbeiführen. Concept Mapping stellt eine Alternative zu klassischen Diagnoseverfahren in der Schule dar.

Concept Maps können außerdem in Schule und Hochschule in Lehr- und Lernsituationen und in den Hochschuldidaktiken als Forschungsmethode und Forschungsobjekt eingesetzt werden (vgl. u. a. Haugwitz & Sandmann, 2009; Stracke, 2004). Der letzte Punkt wird im folgenden Abschnitt näher erläutert.

2.2.3 Forschungsergebnisse zum Einsatz mit und zur Qualität von Concept Mapping

Concept Maps sind in den 80-er Jahren selbst zum Forschungsgegenstand geworden. Basierend auf der Theorie Ausubels des ‚meaningful learning' (1960) und der Annahme, dass, ausgehend von den Befunden zur Concept Map-Forschung in der Biologie, Wissen hierarchisch aufgebaut ist, wurde durch Joseph Novak und Kollegen das Concept Mapping für den englischsprachigen Raum als neues Forschungsfeld etabliert (Novak & Gowin, 1984; vgl. Stracke, 2004). Allerdings muss darauf hingewiesen werden, dass parallel dazu mit der sogenannten Heidelberger Struktur-Lege-Technik von Scheele und Groeben (1984) das Concept Mapping unter einem anderen Namen in den

deutschsprachigen Forschungsraum eingeführt wurde. Die Methodik wurde allerdings erst durch Novak unter dem Begriff Concept Mapping weiter theoretisch fundiert und bekannt.

Die Forschung mit und über Concept Mapping kann mittlerweile auf eine beträchtliche Anzahl von Publikationen zurückblicken (vgl. Hattie, 2009; Nesbit & Adesope, 2006; Horton, McConney, Gallo, Woods, Senn & Hamelin, 1993). Die Studien beschäftigen sich u. a. mit Concept Mapping (CM) und Lernerfolg (u. a. Hucke & Fischer, 2003), CM als Lernhilfen für kollaboratives Lernen (u. a. Haugwitz, 2009; Patterson, Dansereau & Newbern, 1992), CM als Lernstrategie (u. a. Wahser, 2007; Renkl & Nückles, 2006) und CM im Vergleich zum Lernerfolg anderer Lernmethoden (vgl. Metaanalyse von Hattie, 2009). Weiter werden die Gütekriterien wie die Objektivität, die Reliabilität und die Validität von Concept Maps (u. a. Conradty & Bogner, 2012; Ingeç, 2009; Ruiz-Primo, Schultz, Li & Shavelson, 2001; McClure, Sonak & Suen, 1999), Concept Maps als Planungshilfen (u. a. Trochim, 1989; Starr & Krajik, 1990), die Softwareentwicklung für die Erstellung und Auswertung von Concept Maps (z. B. Ifenthaler, 2010) und allgemeiner von Netzwerken (u. a. Fürstenau & Trojahner, 2005; Handcock, Hunter, Butts, Goodreau & Morris, 2008), die graphentheoretische Auswertung (u. a. Borgatti & Everett, 2006; Bonato, 1990; Mavanga, 2001) und Experten- und Novizen-Concept Maps (u. a. Friege & Lind, 2000) zum Gegenstand von Untersuchungen gemacht. Die Studien werden überwiegend als experimentelle Designs angelegt. In der empirischen Forschungsliteratur scheint Concept Mapping als Instruktionshilfe und zur Beurteilung von Schülerkognition eine besondere Bedeutung zu besitzen (vgl. Ruiz-Primo & Shavelson, 1996).

Bei Betrachtung der genannten Studien wird deutlich, dass es viele verschiedene Möglichkeiten gibt, Concept Maps einzusetzen und zu erstellen. Das Fehlen fester Regeln und Vorgaben, beispielsweise bei der Konstruktion von Concept Maps, befördert diese Vielfalt. Ruiz-Primo und Shavelson (1996) haben die drei Kategorien *Aufgabenformat, Antwortformat* und *Bewertungsformat* eingeführt, um Concept Maps zur Beurteilung systematisch beschreiben zu können.

Unter *Aufgabenformat* verstehen die Autoren beispielsweise Aufgabenstellungen, in denen die Concept Map-Ersteller eine bereits vorgefertigte Concept Map analog zu einer Lückentextaufgabe ergänzen sollen (vgl. u. a. Studie von Anderson & Huang, 1989). Weitere Variationen werden in Tabelle 2.1 dargestellt.

Tabelle 2.1. Beispiele verschiedener Concept Map-Aufgabenformate in der Forschung.

	Beschreibung	**Quelle**
Aufgabenformat	Freie Aufgabe, vorgegebene Anzahl an Begriffen	u. a. Barenholz & Tamir, 1992
	Vorgabe einer hierarchischen oder netzartigen Map	u. a. Markham, Mintzes & Jones, 1994; Novak, Gowin & Johansen, 1983
	Vorgegebene Relationen	u. a. McClure & Bell, 1990; Anderson & Huang, 1989
	Weitere Begriffe dürfen ergänzt werden	u. a. Hucke & Fischer, 2000
	Integrative Map: Alltagsbegriffe, Fachbegriffe	u. a. Sumfleth & Tiemann, 2000

Bemerkung: Diese Tabelle ist angelehnt an Ruiz-Primo und Shavelson, 1996.

Die Schwierigkeit der Erstellung der Map wird über die Vorgaben geregelt. Striktere Vorgaben erleichtern die Konstruktion, eine offene Aufgabenstellung erhöht die Schwierigkeit. Kombinationen aus diesen Vorgaben sind denkbar, um beispielsweise eine Integration von Alltagswissen und Fachwissen durch die Aufgabenstellung zu erzwingen (vgl. Sumfleth & Tiemann, 2000). Die Aufgabenstellung richtet sich danach, welche Art von Produkt erzeugt und welches diagnostische Ziel verfolgt werden soll.

Die Bearbeitung der Concept Map wird bei Ruiz-Primo und Shavelson (1996) unter der Kategorie *Antwortformat* beschrieben. Hierbei können beispielsweise auf einem DIN A3-Blatt die Maps per Bleistift oder im Sinne eines Multimedia-Learning-Ansatzes durch ein PC-Programm erstellt werden (vgl. u. a. Studien von Acton, Johnson & Goldsmith, 1994; Fisher, 1990; McClure & Bell, 1990; Beyerbach & Smith, 1990). Allerdings erfordert in beiden Fällen die Erstellung der Map Erfahrung mit dem jeweiligen Medium (vgl. u. a. Plötzner, Leuders & Wichert, 2009).

Nach Ruiz-Primo und Shavelson (1996) gibt es drei Möglichkeiten Concept Maps *auszuwerten*: Durch die Bewertung einzelner Komponenten der individuellen Map, durch Vergleich der Map mit einer Expertenmap bzw. Beispielmap und durch die Kombination aus individueller Bewertung einzelner Mapkomponenten und des Vergleichs mit einer Expertenmap. Die Bewertung einzelner Komponenten kann inhaltlich oder auf Basis graphentheoretisch struktureller Ansätze betrachtet werden, bei denen beispielsweise die Anzahl der genutzten Begriffe gezählt wird oder der Durchmesser und die Dichte der Concept Map berechnet werden (vgl. u. a. Bonato, 1990, Ifenthaler, 2010). Es wird derzeit

uneinheitlich gesehen, inwiefern graphentheoretische Auswertungen Aufschluss über den inhaltlichen Gehalt einer Concept Map geben.

Für dieses Projekt, das u.a. ein schulpraktikables Aufgabenformat entwickeln möchte, wird eine Kombination aus zwei Aufgabenbestandteilen genutzt. Um ein Mindestmaß an Concept Map für eine Diagnose generieren zu können, werden den Lernern Fachbegriffe vorgegeben (Bestandteil 1) und um einen Bezug zum Alltagswissen der Lerner zu erhalten, Bilder von physikalischen Alltagssituationen (Bestandteil 2) vorgelegt (vgl. u. a. Tiemann, 1999). Die Lerner erstellen die Concept Maps mit Papier und Bleistift. Die Maps werden über ein ganzheitliches Verfahren bewertet[2].

Ein anderer, auch für diese Arbeit wichtiger Aspekt zur Beurteilung von Concept Maps, sind die *Gütekriterien*. Die folgenden Gütekriterien sind in der Diskussion:

- Ist eine Concept Map-Erstellung unabhängig von dem Lerner? (Objektivität)
- Wie genau und zuverlässig misst eine Concept Map eine Fähigkeit? (Reliabilität)
- Misst eine Concept Map genau die Fähigkeit, die gemessen werden soll? (Validität)

Die Forschungslandschaft zeigt ein breites Bild an Studien, die die Reliabilität und Validität untersucht haben (vgl. Ruiz-Primo & Shavelson, 1996). Als Grundtenor kann festgehalten werden, dass die Aufgabenstellung und das Concept Map-Bewertungsformat den Weg zur Berechnung der Güte einer Concept Map bestimmen (vgl. u. a. Ingeç, 2009; McClure, Sonak & Suen, 1999).

Die Mehrheit der Studien, die Concept Maps nutzen, machen wenige Aussagen zur Reliabilität des genutzten Concept Map-Verfahrens (Ruiz-Primo & Shavelson, 1996). Der Weg, wie die Reliabilität berechnet wird, ist uneinheitlich. In den meisten Fällen wird von Interraterreliabilitäten (oder Mehrfachbeurteilungen) von Concept Maps gesprochen, indem die Urteile mehrerer Rater zu verschiedenen Komponenten einer Concept Map verglichen werden. Bei Lay-Dopyera und Beyerbach (1983) wird die Übereinstimmung der Beurteiler unter anderem in der Feststellung der Anzahl der genutzten Begriffe ermittelt. Sie berichten für dieses Beispiel einen Interraterkoeffizienten nach Pearson von $r = 0.99$ (Lay-Dopyera & Beyerbach, 1983). Einige Studien geben nicht an, wie die Reliabilität bestimmt wird (vgl. u. a. Anderson & Huang, 1989; Fisher, 1990). Diejenigen

[2] Die Instrumente werden im Kapitel 4.1.3 detailliert beschrieben.

Studien, die über Reliabilitäten berichten, nutzen die Stabilität der Beurteilung der Concept Maps als Reliabilitätsmaß (vgl. Metaanalyse von Ruiz-Primo & Shavelson, 1996). Allerdings muss erwähnt werden, dass die akzeptablen Interraterreliabilitäten von der Concept Map-Komponente (z. B. das Auszählen genutzter Begriffe) abhängig sind, die beurteilt werden soll. Die Reliabilität wird in einigen Studien dadurch verbessert, dass die Beurteiler nach strikten Vorgaben in der Beurteilung geschult wurden (vgl. Schecker & Klieme, 2000). Wie Lehrerinnen und Lehrer Concept Maps beurteilen und wie hoch ihre Reliabilitäten ausfallen, wurde bisher nur vereinzelt in den Fokus genommen (vgl. Lomask, Baron, Greig & Harrison, 1992). Dies deutet darauf, dass die Studien vermehrt abseits vom praktischen Einsatz in Schulen durchgeführt wurden.

Die Studien, die über Validitäten berichten, weisen ein breites Spektrum an Validitätswerten auf. In vielen Fällen werden Concept Maps konvergent (bzw. konkurrent) und divergent gegen ein anderes externes Instrument eingesetzt (vgl. u. a. Ingeç, 2009; McClure, Sonak & Suen, 1999; Ruiz-Primo, Schultz, Li & Shavelson, 2001; Ruiz-Primo, 2000; Schecker & Klieme, 2000). Korrelationen geben an, inwiefern beide Verfahren das gleiche Merkmal messen und wie hoch der Zusammenhang ist. Die Forschungslage zeigt, dass von einer bestimmten Validität nicht gesprochen werden kann. Jede Studie für sich erfasst durch das genutzte Aufgaben- und Bewertungsformat Komponenten von Wissen, die andere Concept Map-Aufgaben- und Bewertungsformate nicht erfassen können. Die Validität ist deshalb nicht unabhängig vom Design der jeweiligen Studie zu diskutieren (vgl. Fischler & Peuckert, 2000; McClure, Sonak & Suen, 1999). Inhaltlich eng gefasste Concept Map-Aufgabenformate, wie das Ausfüllen einer Lücken-Concept Map, korrelieren mit klassischen Leistungstestaufgaben sehr hoch (vgl. u. a. Studie von Anderson & Huang, 1989). Die Ergänzungen der Lücken werden mit ‚richtig-falsch' beurteilt, ebenso die Leistungstestaufgaben, sodass das Bewertungsformat dasselbe ist. Die konvergente Validität von offenen Concept Map-Aufgabenformaten (z. B. sollen beliebig viele Begriffe einer vorgegebenen Wortliste genutzt werden) korreliert mit einem Leistungstest sehr niedrig. Einige Studien berichten über nicht signifikante Korrelationen. Dies liegt unter anderem an unterschiedlichen Bewertungen der zu vergleichenden Instrumente. Während der Leistungstest dichotom bewertet wird, kann die Concept Map holistisch über ein Rating betrachtet werden oder durch die Anzahl von richtigen und falschen Propositionen (vgl. u. a. Studie von McClure, Sonak & Suen, 1999). Zusätzlich können die niedrigen Validitäten dadurch erklärt werden, dass die verglichenen Instrumente unterschiedliche Fähigkeiten messen. Die verschiedenen Studien mit

unterschiedlichen Bewertungsformaten entsprechen den Erwartungen, dass ein Leistungstest das reine Fachwissen testet und bereits von seiner Anlage her nicht das gleiche messen kann, wie ein offenes Concept Map-Aufgabenformat, das nur teilweise Fachwissen erfasst. Dadurch können je nach Studiendesign keine hohen Validitäten erwartet werden. Wie Fischler und Peuckert (2000) in ihrer Übersicht beschreiben:

„ […] Eine generelle Aussage kann es aufgrund der vielfältigen Möglichkeiten für die Gestaltung und Bewertung von Concept Maps auch gar nicht geben. Einige Untersuchungen haben signifikante Korrelationen zwischen Concept Map-Bewertungen und aus anderen Verfahren gewonnenen Wissensindikatoren festgestellt, wobei sichtbar wird, dass ein Bewertungsschema, das sich auf die Prüfung der Richtigkeit der angegebenen Relationen konzentriert, also sich mehr an der inhaltlichen Güte als an topografischen Strukturmerkmalen orientiert, am ehesten zu annehmbaren Korrelationen gelangt (Rice, Ryan & Samson, 1998, McClure et al., 1999). […] “ (Fischler & Peuckert, 2000, 19). Die Spanne der berichteten konvergenten und divergenten Validitäten reicht von r = -0.02 (Novak, Gowin & Johansen, 1983) bis r = 0.82 (Ruiz-Primo, Schultz, Li & Shavelson, 2001).

Concept Mapping wird von der Forschung als Diagnoseinstrument für die Wissensstrukturen von Schülergruppen, Studierenden und Erwachsenen eingesetzt. In vielen Fällen sind es die Forscher, die die Probanden in das Concept Mapping Verfahren einführen und die Concept Maps evaluieren. Die überwiegend englischsprachigen Studien zum Concept Mapping sehen die Lehrpersonen ausschließlich als Concept Map-Ersteller oder Concept Map-Beurteiler. Darüber wie Lehrpersonen den Nutzen von Concept Maps hinsichtlich Praxistauglichkeit und Diagnose einschätzen, wird bislang nicht berichtet. Dies lässt darauf schließen, dass Concept Mapping als Diagnoseinstrument in der Praxis von Lehrerinnen und Lehrern nicht genutzt wird.

2.2.4 Einsatz von Concept Maps als Diagnoseinstrument im Physikunterricht

Die abschließende Betrachtung des Kapitels über Concept Maps soll die wesentlichen Aspekte hervorheben, die Lehrpersonen beim Einsatz von Concept Maps im Physikunterricht berücksichtigen sollten. Concept Maps werden in der Forschung als Diagnoseinstrument eingesetzt. In der Schule werden sie in Deutschland bislang wenig genutzt.

Eine Lehrperson sollte sich im Vorfeld fragen, warum die Schülerinnen und Schüler Concept Maps erstellen sollen. Dies bedeutet, dass die Lehrperson das *diagnostische Ziel* und den Zweck (z. B. für das Erfassen des Vorwissens zu Beginn einer Unterrichtsreihe) festlegen muss.

Entscheidet sich die Lehrperson für den Einsatz von Concept Maps, muss sie zunächst darauf achten, dass die *Aufgabenstellung* für die Schülerinnen und Schüler nicht zu komplex ist (vgl. u. a. Jüngst & Strittmatter, 1995, McClure, Sonak & Suen, 1999). Damit der diagnostische Einsatz von Concept Maps für die Lehrkraft möglich wird, müssen die Maps nicht nur inhaltlich gehaltvoll, sondern ebenfalls zeitökonomisch *auswertbar* sein (vgl. Jüngst & Strittmatter, 1995). Die Auswertung einer Concept Map kann, je nach Zeit, die die Lehrkraft investieren möchte, unterschiedlich aussehen. Es ist denkbar, dass die Lehrkraft ohne Regeln auf die Concept Map blickt und versucht die Wissensstrukturen, z. B. hinsichtlich der für den Unterricht relevanten Schülervorstellungen für eine schnelle Diagnose zu erfassen. Wenn die Map systematisch unter bestimmten Gesichtspunkten (z. B. ob eine bestimmte Verknüpfung dargestellt wurde) werden soll, kann die Lehrperson sich zuvor schriftlich einen Erwartungshorizont erstellen. Die Lehrkraft kann festlegen, welche korrekten Propositionen (Begriff-beschrifteter Pfeil-Begriff) sie in den Concept Maps der Schülerinnen und Schüler erwartet oder welche zentralen Begriffe sie verlangt.

Concept Maps können, je nach Aufgabenstellung, in ihrer Darstellung komplex werden, sodass ein *„Lesen“* der Map für Lehrerinnen und Lehrer schwierig werden kann. Es ist beispielsweise nicht klar, wo begonnen werden soll, um die Map zu lesen. Basierend auf den Ergebnissen dieser Arbeit zur Zeitökonomie einer Concept Map-Bewertung (siehe Kapitel 5.1 und 5.2 Ergebnisse zur Zeitökonomie) kann angenommen werden, dass mit zunehmender Anzahl von Concept Map-Bewertungen das Lesen der Map schneller gelingt und kürzer und einfacher wird. Beim Lesen von Concept Maps scheint also ein Übungseffekt einzutreten.

Eine Lehrperson muss ebenfalls eine Entscheidung über das *Medium* treffen, in dem die Concept Maps erstellt werden sollen. Für den schnellen Einsatz bieten sich Papier und Bleistift an. Sie sind kostengünstig und Schülerinnen und Schüler nutzen täglich Papier und Stifte. Concept Maps, die am Computer erstellt werden, sind eine Alternative. Der Umgang mit den für diesen Zweck entwickelten Programmen muss aber geübt werden (vgl. Nückles, Gurlitt, Pabst & Renkl, 2004).

Der *Zeitpunkt*, wann die Lehrkraft die Schülerinnen und Schüler auffordert, eine Concept Map zu erstellen, hängt, wie das Aufgabenformat, vom *diagnostischen Ziel* ab. Der Einsatzzeitpunkt kann von der Lehrperson frei gewählt werden. Im unterrichtlichen Verlauf bietet sich die Concept Map-Erstellung für eine Vorwissensabfrage, unterrichtsbegleitende Erstellung oder als Abschluss einer Unterrichtseinheit an. Der Lehrperson wird, unabhängig davon, wann sie Concept Maps erstellen lässt, jederzeit eine Status- und Prozessdiagnostik über die eigenen Schülerinnen und Schüler ermöglicht. Wird beispielsweise eine Concept Map einmalig in einer Unterrichtsreihe erstellt, kann dies Aufschluss über die aktuelle Wissensstruktur eines Lerners geben. Werden hingegen mehrere Concept Maps während einer Unterrichtsreihe erstellt, kann die Entwicklung der Wissensstruktur diagnostiziert werden.

Wenn die Schülerinnen und Schüler Concept Maps erstellen sollen, muss zuvor ein *Training* durchgeführt werden, um eine gewisse Vertrautheit/Routine in der Nutzung dieser Methodik für die Schülerinnen und Schüler entstehen zu lassen (vgl. u. a. Jüngst & Strittmatter, 1995, Schau & Mattern, 1997). Der empfohlene Zeitaufwand für eine Einübung des Verfahrens wird in der Literatur unterschiedlich angesetzt. Programme, die dieses Training fördern, wie die von Sumfleth, Neuroth und Leutner (2010) bedürfen einer Übungsperiode von 60 Minuten. Im Rahmen der hier vorgestellten Forschungsarbeit hat sich gezeigt, dass bereits Trainingsstunden von 45 Minuten ausreichend sind.

Entschließt sich eine Lehrkraft Concept Maps erstellen zu lassen, muss den Schülerinnen und Schülern erklärt werden, ob die Erstellung der Map mit einer Leistungsabfrage verbunden ist oder ob sie eine rein diagnostische Funktion besitzt und der Lehrperson zur Adaption ihres Unterrichts dient.

2.3 Zusammenfassung

Das Arbeitsfeld von Lehrkräften aller Unterrichtsfächer umschließt neben Unterrichten, Erziehen und Innovieren den Bereich des *Beurteilens* im Rahmen einer pädagogisch-psychologischen Diagnostik. Diagnostik soll in Schulen u. a. betrieben werden, um Lernzustände von Schülerinnen und Schülern zu erfassen, mit dem Ziel, eine Passung des Unterrichts auf die Bedürfnisse der Schülerinnen und Schüler vorzunehmen. Eine Diagnose im Klassenzimmer setzt allerdings ein fundiertes Wissen über *diagnostische*

Theorien voraus, die von einer individuellen Statusdiagnostik bis zu einer Prozessdiagnostik reichen können. Ebenfalls muss die Lehrkraft über Wissen zu *diagnostischen Methoden,* wie der Verhaltensbeobachtung, Befragungsmethoden oder Testmethoden und über das Wissen über ihren *Einsatz* verfügen. Unter anderem ist dieses Wissen in der *diagnostischen Kompetenz* eingebettet.

Es hat sich gezeigt, dass diagnostische Kompetenz, gemessen als Diagnosegenauigkeit (Schrader & Helmke, 1987), bei den untersuchten Lehrkräften des deutschsprachigen Raumes unterschiedlich stark ausgeprägt ist. Es gibt neben der von Helmke und Schrader vorgeschlagenen *Diagnosegenauigkeit* noch kein weiteres Verfahren, Diagnosekompetenz zu operationalisieren und messbar zu machen. Aus Gründen der Vergleichbarkeit der Messungen wird in dieser Arbeit deshalb die Diagnosekompetenz als Maß für die Diagnosegenauigkeit betrachtet. Das Maß Diagnosegenauigkeit wird aus der *Rangordnungskomponente,* der *Niveaukomponente* und der *Streuungskomponente* gebildet. Die Rangordnungskomponente lässt sich als Rangkorrelation zwischen der Leistung von Schülerinnen und Schülern, z. B. in einem Wissenstest, und der von den Lehrkräften eingeschätzten Leistung abbilden. Die Niveaukomponente gibt an, inwiefern eine Lehrkraft Schülerinnen und Schüler über- bzw. unterschätzt. Die Streuungskomponente beschreibt die Streuung des Lehrerurteils im Verhältnis zur Leistungsstreuung der Schülerinnen und Schüler, die mit einem anderen Instrument gemessen wurde (z. B. mit einem Test).

Lehrkräfte können nicht nur Schülerinnen und Schüler einschätzen, sondern ebenfalls Lern- oder Leistungstestaufgaben in ihrer Schwierigkeit beurteilen. Dadurch können Aussagen getroffen werden, wie gut Lehrerinnen und Lehrer Aufgabenschwierigkeiten einschätzen können.

Momentan werden weitere Ansätze verfolgt, die Diagnosekompetenz zu operationalisieren, die allerdings noch in einem Entwicklungsprozess sind. Wenn Lehrkräfte ihre Diagnosegenauigkeit angemessen ausbilden sollen, setzt dies unter anderem voraus, dass sie geeignete *Diagnoseinstrumente* entwickeln können bzw. ihnen bereits evaluierte Diagnoseinstrumente zur Verfügung stehen. Für Physiklehrkräfte ist der Umfang an erreichbaren Diagnoseinstrumenten bisher noch begrenzt. Dies wird zum Anlass genommen mit Concept Maps ein schulpraktisches Diagnoseinstrument zu erstellen.

Nachdem Lerner das Verfahren Concept Mapping erlernt haben, ermöglichen es *Concept Maps*, die Wissensstrukturen und Konzepte des Lerners zu erfassen. Dabei

können Schülervorstellungen deutlich werden, die es der Lehrkraft ermöglichen, im Unterrichtsverlauf angemessen auf diese Vorstellungen zu reagieren.

Vom wissenschaftlichen Standpunkt wird seit jeher viel mit und über Concept Maps geforscht. Concept Maps werden beispielsweise in der Forschung als Diagnoseinstrument eingesetzt. Neben dieser Einsatzoption existieren viele verschiedene Befunde beispielsweise zu den Gütekriterien von Concept Maps. Eindeutige Aussagen zur konvergenten und divergenten Validität können nicht getroffen werden, da diese maßgeblich von der entsprechenden Studie, die die Validität berichtet, abhängen. Die verschiedenen Studien setzen unterschiedliche Aufgabenformate und Bewertungsformate von Concept Maps ein. Durch diese Faktoren wird die Höhe der Validität bestimmt. Zusätzlich bleibt offen, welche Anteile an Wissen und Kompetenzen Concept Maps messen können. In vielen Studien wird berichtet, dass Concept Maps Teile von Wissen erfassen, die mit anderen Verfahren verborgen bleiben.

Ausgehend von der Annahme einer entwickelbaren Diagnosegenauigkeit von Lehrkräften und der Tatsache, dass mit Concept Maps ein Diagnoseinstrument gegeben ist, schließt der theoretische Rahmen dieser Arbeit mit der zusammenfassenden Abbildung 2.4. Im nachfolgenden Kapitel werden die leitenden Forschungsfragen und Hypothesen dieser Arbeit abgeleitet.

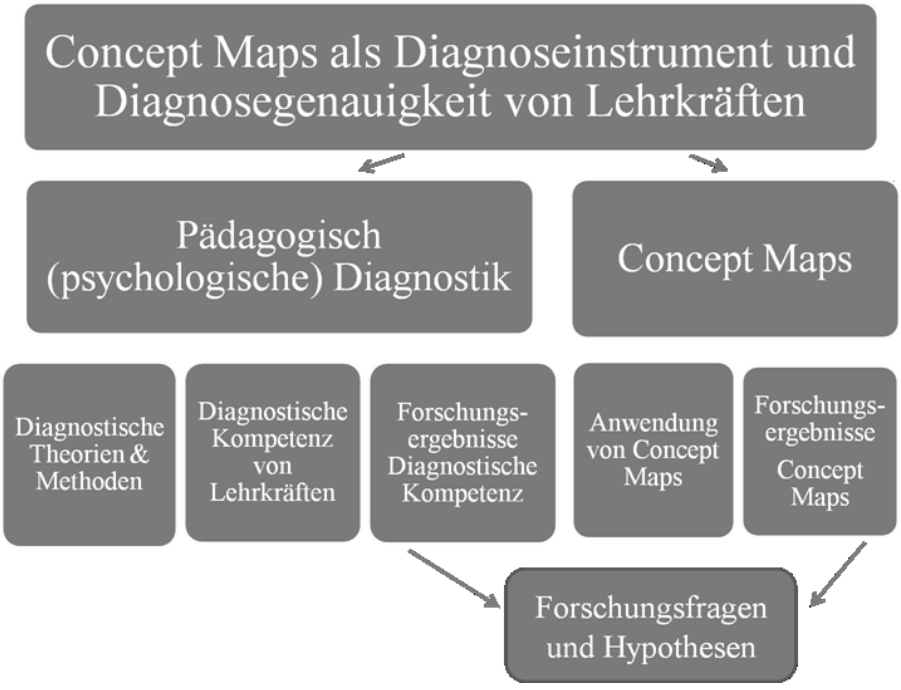


Abbildung 2.4. Zusammenfassende Übersicht des theoretischen Rahmens dieser Arbeit.

3 Ziele, Forschungsfragen und Hypothesen

Aus dem theoretischen Rahmen dieser Arbeit sind drei zentrale Elemente für die Forschungsarbeit ausschlaggebend:

- die Annahme einer nicht optimalen Diagnosegenauigkeit von Lehrkräften,
- die bislang noch in geringer Anzahl existierenden Diagnoseinstrumente für Physiklehrkräfte und
- die Möglichkeit, Concept Maps als Diagnoseinstrument zu nutzen.

Für den Einsatz im Physikunterricht soll ein angemessenes Concept Map-Aufgabenformat mit einer zeitökonomischen Bewertung entwickelt werden, das didaktisch relevante, diagnostische Informationen erzeugt und den wissenschaftlichen Standards genügt. Es muss also für den Praxiseinsatz ein Aufgabenformat erstellt werden, das reliabel und valide ist, bevor es den Physiklehrkräften zur Erprobung gegeben werden kann.

Diese Forschungsarbeit gliedert sich in zwei Studien, mit denen, aufeinander aufbauend, zunächst zwei Concept Map-Aufgabenformate und ein Bewertungsformat ausgewählt werden und deren *Validität* untersucht wird (Studie 1). Anschließend wird ein Aufgabenformat gewählt und mit diesem die *Diagnosegenauigkeit* von Physiklehrkräften gemessen (Studie 2).

Im Vorfeld der Studie 1 werden sieben Concept Map-Aufgabenformate explorativ entwickelt und von Schülerinnen und Schülern bearbeitet. Aus diesen Aufgabenformaten werden zwei für die Studie 1 ausgewählt. Die detailliertere Beschreibung dieser Vorstudie erfolgt in Kapitel 4.1.3 Beschreibung der Instrumente.

In Studie 1 werden aus *zwei* unterschiedlichen Concept Map-Aufgabenformaten und einem Bewertungsformat ein Aufgabenformat für die Studie 2 nach Validität und Reliabilität ausgewählt. Dazu werden beide Aufgabenformate jeweils mit dem einen Bewertungsformat konvergent gegen ein standardisiertes Testinstrument validiert, von dem ausgegangen werden kann, dass es Physikkompetenzen erfasst. Die Frage der ersten Studie lautet entsprechend:

> *FF 1. Welcher Zusammenhang besteht zwischen Aufgabenformat und Bewertungsformat von Concept Maps und den in einem Kompetenztest gemessenen Schülerkompetenzen?*

Für diese Studie werden aus der Theorie begründet zwei Concept Map-Aufgabenformate entwickelt[3]:

1. Aufgabenformat A: Schülerinnen und Schüler erhalten eine Wortliste zum Basiskonzept Energie und sollen eine Concept Map erstellen. Diese Aufgabenstellung ist durch die vorgegebenen Fachbegriffe auf fachsprachlicher Ebene einzuordnen.

2. Aufgabenformat B: Schülerinnen und Schüler erhalten drei Bilder zu physikalischen Situationen im Bereich Energie und sollen auf dieser Basis eine Concept Map generieren. Anschließend sollen sie ihre Concept Map um die Begriffe aus Aufgabenformat A ergänzen. Dieses Aufgabenformat ist sowohl alltagssprachlich, anwendungsorientiert (erster Aufgabenteil) als auch fachsprachlich (zweiter Aufgabenteil) ausgerichtet.

Die anschließende Bewertung der Concept Maps beider Aufgabenformate erfolgt jeweils über den Concept Map-Beurteilungsbogen, der in Anlehnung an Diagnosebögen aus den Fächern Deutsch und Mathematik für diese Studie entwickelt wurde.

Die Forschungsergebnisse zum Concept Mapping lassen bereits darauf schließen, dass klassische Testinstrumente wie Multiple-Choice-Tests nicht vollends das gleiche Konstrukt messen wie Concept Maps (vgl. u. a. Studien von Anderson & Huang (1989); Ingeç (2009)). In Anlehnung an die Thematik, welche Anteile an Wissen und Kompetenzen Concept Maps messen (vgl. Abschnitt 2.2.3), wird begründet angenommen, dass mit den entwickelten Aufgabenformaten und dem Bewertungsformat Kompetenzen, wie sie im Kompetenztest getestet werden, partiell abgebildet werden können. Damit kann die Frage, welche Fähigkeit Concept Maps nicht messen, weiter erschlossen werden. Die Hypothesen, die aus FF 1 resultieren, gliedern sich in zwei Bereiche.

Mit Hypothese H1.1 wird die erwartete konvergente Validität beschrieben:

H1.1 Es besteht eine positive Korrelation im unteren Bereich zwischen Bewertung der Concept Maps über Beurteilungsbogen und Kompetenztest.

Die bestehenden Forschungsbefunde lassen begründet annehmen, dass mittlere Korrelationen zwischen geschlossenen Tests (hier der Kompetenztest) und Concept Map-Aufgabenformaten mit offenem Antwortcharakter zu erwarten sind (vgl. Review Ruiz-Primo & Shavelson, 1996).

Hypothese H1.2 differenziert, welches der beiden entwickelten Aufgabenformate mit dem externen Validierungsinstrument stärker korreliert:

[3] An dieser Stelle wird für eine detaillierte Beschreibung des Instrumentenentwicklungsprozesses auf Kapitel 4.1.3 verwiesen. Um Begründungszusammenhänge zu verdeutlichen, werden an dieser Stelle die Instrumente kurz erläutert.

H1.2 Es besteht eine höhere Korrelation zwischen Aufgabenformat A mit dem Kompetenztest als zwischen Aufgabenformat B und dem Kompetenztest.

Es wird davon ausgegangen, dass Aufgabenformat A mit dem Kompetenztest höher korreliert, da beide Instrumente in ihrem Typus ähnlich sind. Beide Instrumente sind auf einer fachsprachlich inhaltlichen Ebene anzusiedeln (vgl. Schecker & Klieme, 2000). In Aufgabenformat B müssen die Schülerinnen und Schüler die Concept Map ebenfalls auf fachsprachlicher Ebene erstellen. Allerdings entsteht durch die erste Phase mit den Bildern zusätzlich ein anwendungsorientierter, auf Alltagserfahrungen orientierter Bezug; die Ergebnisse dieser Concept Maps sollten niedriger mit dem Kompetenztest korrelieren.

Auf Basis der Ergebnisse der Studie 1 soll für Studie 2 eine Entscheidung zu Gunsten eines Aufgabenformats getroffen werden, das den Physiklehrerinnen und Physiklehrern als Diagnoseinstrument angeboten wird. Das Bewertungsformat in Form des bereits in Studie 1 genutzten Concept Map-Beurteilungsbogens wird den Lehrkräften ebenfalls angeboten, sodass untersucht werden kann, inwiefern das Concept Map-Aufgabenformat und das Bewertungsformat Einfluss auf die Diagnosegenauigkeit haben. Es können auf dieser Basis Aussagen getroffen werden, inwieweit das Aufgabenformat und Bewertungsformat zur Diagnose geeignet sind. Die Forschungsfrage 2 lautet:

FF 2. Inwiefern sind Concept Maps ein geeignetes Instrument für Lehrerinnen und Lehrer zur Diagnose von Schülerkompetenzen im Physikunterricht?

Die Eignung dieses Diagnoseinstrumentes soll über die Diagnosegenauigkeit der Physiklehrkräfte gemessen werden (vgl. Abschnitt 2.1.2). Ausgehend von der Theorie zur Operationalisierung der Diagnosegenauigkeit wird in Hypothese 2.1 generell davon ausgegangen:

H 2.1 Physiklehrerinnen und Physiklehrer sind in der Lage, durch die Diagnose ihrer Schülerinnen und Schüler mit Concept Maps (CM) und dem Bewertungsformat Concept Map-Beurteilungsbögen (CM-BB) eine Rangordnung zu erstellen, die der Rangordnung eines Kompetenztests ähnlich ist.

Die Höhe dieser Rangordnungsübereinstimmung wird in Hypothese 2.2 detaillierter beschrieben:

H 2.2 Die Rangordnungsübereinstimmung (Diagnosegenauigkeit) gemessen als Rangkorrelation ist am höchsten, wenn beide Instrumente (CM & CM-BB) zusammen eingesetzt werden.

Es wird davon ausgegangen, dass die Lehrperson ihre Schülerinnen und Schüler am genausten einschätzen kann, wenn beide Instrumente kombiniert eingesetzt werden.

Bevor das Kapitel Methoden und Design die Anlage dieses Projektes erklärt, erfolgt im nachfolgenden Kapitel eine Einordnung des Projektes in das Forschungsgenre.

Exkurs: Feldstudien

Empirische Studien können je nach Zielsetzung unterschiedlich angelegt und strukturiert sein. Im Allgemeinen kann zwischen e*xperimentellen* und *quasiexperimentellen Untersuchungen* unterschieden werden. Eine weitere Betrachtung in *Labor-* und *Felduntersuchungen* erlaubt zusätzlich eine Feinaufgliederung in vier Untersuchungsvariationen: experimentelle Laboruntersuchung, experimentelle Felduntersuchung, quasiexperimentelle Laboruntersuchung und quasiexperimentelle Felduntersuchung (vgl. Bortz & Döring, 2006; Sedlmeier & Renkewitz, 2008).

Experimentelle Untersuchungen zeichnen sich dadurch aus, dass die Teilnehmerinnen und Teilnehmer einer Studie randomisiert auf die Versuchsgruppen aufgeteilt werden. Quasiexperimentelle Untersuchungen hingegen unterscheiden sich von experimentellen Designs, indem mit natürlich existierenden Versuchsgruppen gearbeitet wird. Beispielweise lässt sich ein Physiklehrer mit der 8. Klasse, die er gerade unterrichtet, einem quasiexperimentellen Design zuordnen. Die Auswahl solch einer Gruppe ist nicht zufällig, sondern besteht bereits (vgl. u. a. Sedlmeier & Renkewitz, 2008; Fraenkel, Wallen & Hyun, 2012; Creswell, 2012).

Studien können außerdem nach den Kategorien Laboruntersuchung und Felduntersuchung klassifiziert werden. Der Unterschied besteht darin, dass Laboruntersuchungen in strikten Laborbedingungen durchgeführt werden, in denen Störvariablen kontrolliert bzw. eliminiert werden. Feldstudien finden in einem aktiven Feld (z. B. einer Schule) statt. Sie können einer Vielzahl von Störeffekten unterliegen. Störeffekte sind beispielsweise Baustellengeräusche von der Straße oder eine verminderte Anzahl an teilnehmenden Schülerinnen und Schülern, da an diesem Tag ein Sportfest stattfindet. Feldstudien sind im Vergleich zu Laboruntersuchungen authentischer, da sie direkt im Feld durchgeführt werden und die tatsächlich vor Ort bestehende Realität unverfälscht darstellen (vgl. Bortz & Döring, 2006).

Metaanalysen wie die von Hattie (2009) haben gezeigt, dass teilweise aufwändig geplante Interventionen mit Schülerinnen und Schülern, vom Standpunkt wissenschaftlicher Richtwerte, kleine Effekte aufweisen. Hattie (2009) stellt beispielsweise die Ergebnisse von Interventionsstudien zum problembasierten Lernen dar (Effektstärke d = 0.15). Die Studien können dennoch als Erfolg eingestuft werden, da sie trotz ihrer Vielzahl an nicht kontrollierbaren Parametern überhaupt Effekte aufweisen.

Diese Arbeit lässt sich als quasiexperimentelle Feldstudie einordnen. Alle Elemente dieses Projektes (Studie 1 und Studie 2) sind im natürlichen Raum ‚Schule', von Physiklehrerinnen und Physiklehrern und ihren Physikkursen durchgeführt worden.

4 Methoden, Design und Datenanalyse

In diesem Abschnitt werden die Designs der beiden Studien zur Bearbeitung der Forschungsfragen vorgestellt und die jeweils eingesetzten Instrumente beschrieben. Anschließend werden die konkreten Datenerhebungsschritte der Studien dargestellt und die statistischen Methoden zur Auswertung der erhobenen Daten erläutert.

4.1 Studie 1

4.1.1 Design

In dieser Teilstudie wird die Konstruktvalidität des angewandten Concept Map Verfahrens durch Korrelation mit einem bereits validierten Kompetenztest untersucht. Es wird ein einmaliger Untersuchungszeitpunkt (Querschnittsdesign) gewählt. Die teilnehmenden Schülerinnen und Schüler mehrerer Klassen eines Jahrgangs werden mit einem von zwei möglichen Concept Map-Aufgabenformaten und mit einem Kompetenztest bezüglich ihrer Kompetenz im Basiskonzept Energie getestet.

Die eingesetzten zwei unterschiedlichen Concept Map-Aufgabenformate wurden auf Basis einer explorativen Vorstudie aus sieben Concept Map-Aufgabenformaten ausgewählt (weitere Erläuterungen siehe Abschnitt 4.1.3 Beschreibung der Instrumente).

4.1.2 Stichprobe

An der ersten Studie nehmen 79 Schülerinnen und Schüler aus vier Klassen zweier Gymnasien des neunten Jahrgangs des G8[4] in Nordrhein-Westfalen teil. An dieser Stelle werden keine Angaben zur Geschlechterverteilung, Alter und Intelligenz gemacht, da diese erste Teilstudie nicht die Frage nach Geschlechtereffekten und kognitiven Fähigkeiten verfolgt. Die deskriptiven Statistiken werden in Kapitel 5.1.1 Ergebnisse vorgestellt.

[4]Momentan können Schülerinnen und Schüler in Nordrhein-Westfalen ihr Abitur nach acht Schuljahren oder nach neun Schuljahren machen. Die Abkürzungen lauten daher G8 oder G9. In Nordrhein-Westfalen haben im Sommer 2013 erstmals Schülerinnen und Schüler nach acht und nach neun Jahren Schulzeit gleichzeitig das Abitur erlangt.

4.1.3 Beschreibung der Instrumente

Concept Map-Aufgabenformat

Aus der Theorie abgeleitet, lassen sich Concept Maps unter anderem über ihr Aufgabenformat definieren (Ruiz-Primo & Shavelson, 1996).

Bereits in einer explorativen Studie (Vorstudie), die an dieser Stelle nicht weiter beschrieben wird, wurden sieben verschiedene Aufgabenformate in achten Klassen nordrhein-westfälischer G9-Gymnasien eingesetzt. Die Aufgabenformate orientierten sich an bereits bestehenden Aufgabenformaten der Forschungsliteratur. Das Ziel dieser explorativen Studie war es, geeignete Concept Map-Aufgabenformate für Schülerinnen und Schüler sowie für die Lehrpersonen zu ermitteln. Alle sieben Aufgabenformate haben das Basiskonzept Energie abgefragt. Die Entwicklung der Aufgabenformate orientierte sich am Physikkernlehrplan der Mittelstufe, Physikschulbüchern der Mittelstufe und einer Expertenbefragung hinsichtlich der Begriffsauswahl für das Konzept Energie. Es wurden relevante Begriffe des Basiskonzepts ermittelt, die die Ausgangsbasis einer jeden Concept Map-Aufgabe bildeten. Das Verfahren gewährleistet die inhaltliche Validität der benutzten Begriffe des Verfahrens.

Für das weitere Verfahren werden auf diese Weise zwei aus sieben Aufgabenformaten ausgewählt. Kriterien für die Auswahl der Aufgabenformate sind der Vorbereitungsaufwand für die Lehrkraft, die Zeit für die Durchführung im Unterricht und das Potenzial der Concept Maps als ein Hilfsmittel zur Diagnose der Wissensstrukturen und Konzepte der Schülerinnen und Schüler. Es wird bei dem Einsatz der verschiedenen Aufgabenformate auf eine angemessene kognitive Belastung für die Schülerinnen und Schüler geachtet, durch die die Gefahr einer Über- und Unterbelastung reduziert werden kann (vgl. u. a. Baddeley, 1992; Paas, Tuovinen, Tabbers & Van Gerven, 2003). Dies fordert beispielsweise, dass die Aufgabenstellung für alle Schülerinnen und Schüler sprachlich und inhaltlich verständlich ist und dass die Bearbeitungszeit angemessen ist.

Für Studie 1 werden die Aufgabenformate A und B eingesetzt (vgl. u. a. Tiemann, 1999):

- In Aufgabenformat A erhalten die Schülerinnen und Schüler 21 Begriffe zum Basiskonzept Energie (siehe Anhang A.1). Die Schülerinnen und Schüler sollen aus diesen 21 Begriffen mindestens 10 Begriffe auswählen, mit denen sie eine Concept Map erstellen. Darüber hinaus ist es ihnen freigestellt, weitere Begriffe der Liste zu

benutzen und eigene Begriffe zu ergänzen. Aus den Vorerfahrungen mit den achten Klassen (Vorstudie) wird als Bearbeitungszeit 30 Minuten angesetzt.

- Das Aufgabenformat B enthält zwei Phasen. Die Schülerinnen und Schüler erhalten zunächst drei Bilder zu physikalischen Situationen zum Thema Energie, z. B. ein Kind, das auf einem Trampolin springt (siehe Anhang A.1). Auf Basis dieser Bilder sollen die Schülerinnen und Schüler eine Concept Map zum Thema Energie generieren. Für diese Bearbeitungsphase haben sie 15 Minuten Zeit. Anschließend wechseln die Schülerinnen und Schüler die Stiftfarbe, um den Phasenwechsel in ihrer Concept Map kenntlich zu machen. Sie erhalten wie in Aufgabenformat A die 21-Begriffe-Liste, ebenfalls mit dem Auftrag, 10 Begriffe aus dieser Liste zu wählen und in ihre bereits bestehende Concept Map einzubauen. Weitere Begriffe dürfen genutzt oder ergänzt werden. Diese Phase dauert ebenfalls 15 Minuten.

Beide Aufgabenformate erscheinen nach den Erfahrungen der Vorstudie geeignet zu sein, da sie den Schülerinnen und Schülern einen gewissen Grad an Freiheit in der Nutzung ihnen bekannter Begriffe ermöglichen. In Aufgabenformat A wählen die Schülerinnen und Schüler Begriffe aus der Liste. Aufgabenformat B stellt an die Schülerinnen und Schüler zusätzlich die Anforderung, zunächst mit Alltagswissen und Verständnis der dargestellten Situation eine Concept Map zu erstellen. Später werden diese mit der Liste der Fachtermini verbunden. In beiden Formaten wird durch die Vorgabe, mindestens 10 Begriffe zu nennen, eine minimale Größe der Concept Map angestrebt, um eine Bewertung überhaupt erst zu ermöglichen. Beide Formate geben der Lehrkraft die Möglichkeit, die physikalischen Konzepte der Schülerinnen und Schüler zum Inhaltsbereich Energie zu diagnostizieren. In Abbildung 4.1 wird das Vorgehen zusammengefasst dargestellt.

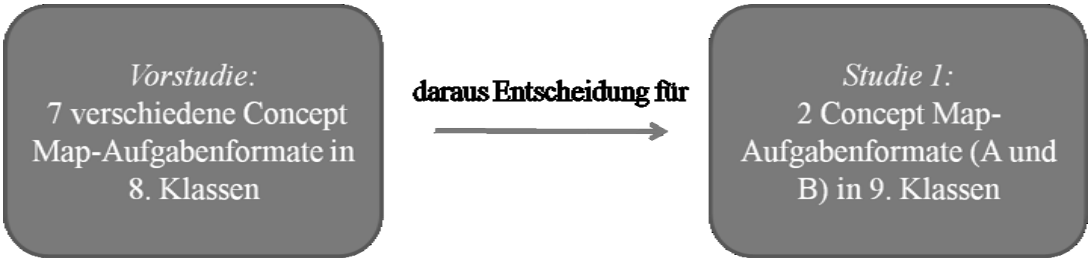


Abbildung 4.1. Vorgehen bei der Auswahl der Concept Map-Aufgabenformate für Studie 1.

Bemerkungen: Die Studie 1 wird mit 9. Klassen durchgeführt. Es konnte mit der Erhöhung der Jahrgangsstufe angenommen werden, dass die Concept Maps eines neunten Jahrgangs inhaltlich umfangreicher sind als die eines achten Jahrgangs und somit eindeutiger zu beurteilen sind.

Concept Map-Beurteilungsbogen als Bewertungsformat

Die Forschungsliteratur beschreibt vielfältige Wege der Auswertung von Concept Maps (u. a. Ruiz-Primo & Shavelson, 1996). Nicht alle Bewertungsformate eignen sich für den Schulalltag. Beispielsweise wird bei einer rein strukturellen Auswertung von Concept Maps, bei der die Anzahl der genutzten Begriffe ausgezählt wird oder der Umfang der Concept Map berechnet wird, die inhaltliche Qualität nicht abgebildet. Diese Art von Auswertung gibt der Lehrkraft keinen Aufschluss über die Begriffe und deren Vernetzungen, über die die Schülerinnen und Schüler zu dem betreffenden Thema verfügen. Die Auszählung von richtigen und falschen Propositionen einer Concept Map eignet sich ebenfalls nur eingeschränkt für eine Diagnose; Begriffe, die in den Concept Maps fehlen, werden z. B. nicht erfasst. Stattdessen werden nur die in der Map existierenden Verknüpfungen bewertet. Es kann lediglich implizit, durch das Fehlen von Verknüpfungen, auf falsche Schülerkonzepte geschlossen werden. Durch das reine Auszählen können zwar mehrere Concept Maps vergleichbar gemacht werden, es ersetzt aber nicht eine zusätzliche inhaltliche Bewertung der Maps.

Ein Bewertungsformat für die Schule muss der Lehrkraft die Möglichkeit geben, den inhaltlichen Gehalt der Concept Maps zu erfassen. Das Verfahren muss außerdem für die Lehrkraft zeitökonomisch und immer wieder einsetzbar sein. Orientiert an diesen Anforderungen und inspiriert durch die breite Materiallage hinsichtlich einer großen Zahl an Diagnosebögen in den Unterrichtsfächern Deutsch und Mathematik (vgl. Paradies, Linser & Greving, 2009), wird ein Concept Map-Beurteilungsbogen zur Auswertung von Schüler-Concept Maps zum Basiskonzept Energie entwickelt. Der Bogen beinhaltet 18 verschiedene Aussagen, die mittels einer Likert-Skala von 0 bis 3 bewertet werden (siehe Kapitel Anhang A.2 Instrumente). Die inhaltliche Validität des Bogens wurde durch einen Vergleich mit dem Curriculum, einschlägigen Schulbüchern und dem benutzten Kompetenztest (siehe unten) sichergestellt. Bei Letzterem wurde darauf geachtet, dass sich der Bogen, ähnlich wie der Kompetenztest, an den theoretischen Annahmen einer Kompetenzentwicklung im Konzept ‚Energie' im Sinne von Liu und McKeough (2005) (vgl. ebenfalls Neumann, Viering & Fischer, 2010) orientiert. Eine Lehrkraft bewertet eine Concept Map durch die Bewertung der Aussagen des Bogens. Die Lehrkraft soll auf einer Likert-Skala von ‚trifft nicht zu (0 Pkt.)' bis ‚trifft völlig zu (3 Pkt.)' feststellen, z. B. ob die Aussage: „Der Schüler/Die Schülerin hat erkannt, dass es verschiedene Energieformen gibt." auf die zu beurteilende Concept Map zutrifft.

Durch dieses Verfahren soll eine inhaltlich systematische Auswertung einer Concept Map ermöglicht werden. Mit dem Bogen kann zusätzlich der Einstieg in die Bewertung einer Concept Map erleichtert werden.

Concept Map-Antwortformat

Nach einer Metaanalyse von Nesbit und Adesope (2006) und einem Übersichtsartikel von Ruiz-Primo und Shavelson (1996) können Concept Maps entweder im papierbasierten oder computergestützten Antwortformat erstellt werden.

Vielfach werden den Concept Map-Erstellern ein Blatt Papier und Stifte zur Verfügung gestellt. Für den Einsatz in der Schule ist dieses Antwortformat preiswert und es stellt keine hohen Anforderungen an die Ressourcen. Das Zeichnen einer Concept Map am Computer erfordert zunächst geeignete Programme. Mittlerweile gibt es eine Vielzahl von Programmen, die dazu genutzt werden können, jedoch teilweise lizenzpflichtig sind (z. B. MaNet, vgl. Eckert, 2000 oder Easy Mapping-Tool, vgl. Nückles, Gurlitt, Pabst & Renkl, 2004). Die Programme, die nicht lizenzpflichtig sind (z. B. CMap Tools des Institute for Human & Machine Cognition (IHMC), 2010), können ohne Kosten auf Schul-PCs installiert werden. Sie sind in der Handhabung allerdings nicht intuitiv und sie schränken den Lerner aus programmtechnischen Gründen in der Erstellung der Concept Map ein (vgl. Nückles, Gurlitt, Pabst & Renkl, 2004). Aus kognitionspsychologischer Sicht können bei der Nutzung von PC-Programmen die Qualität der Einarbeitung (z. B. Zeitfaktor, vgl. Nückles, Gurlitt, Pabst & Renkl, 2004) und die kognitiven Anforderungen während der Concept Map-Erstellung einen Einfluss auf die Qualität der erzeugten Concept Maps haben (zu den kognitiven Anforderungen vgl. u. a. Mayer & Moreno, 2003). Die Komplexität der PC-Programmhandhabung konkurriert offensichtlich mit dem eigentlichen Denk- und Erstellungsprozess der Concept Map.

Ein papierbasiertes Verfahren verringert diesen Konkurrenzeinfluss (vgl. Nückles, Gurlitt, Pabst & Renkl, 2004). Jedoch werden Concept Maps auf Papier schnell unübersichtlich und können, wenn bereits viele Begriffe und Verknüpfungen eingezeichnet wurden, nur mit relativ großem Aufwand geändert werden. Die PC-gestützte Erstellung erleichtet die Änderung von Concept Maps durch ‚anklicken und hin- und herziehen'.

Unter dem Gesichtspunkt der Ressourcen einer Schule wird das papierbasierte Verfahren in dieser Studie eingesetzt.

Kompetenztest (adaptiert)

Um zu prüfen, inwiefern Concept Maps Kompetenzen im Konzept ‚Energie' messen, werden Teile des bereits validierten Kompetenztests von Viering (2012) eingesetzt.

Vierings Test zur Kompetenzentwicklung von Schülerinnen und Schülern im Konzept ‚Energie' orientiert sich an der Theorie von Liu und McKeough (2005). Es wird davon ausgegangen, dass die Schülerinnen und Schüler ihr Verständnis zum Konzept ‚Energie' in vier verschiedenen Stufen (Entwicklungs- oder Kompetenzstufen) entwickeln. Je mehr die Schülerinnen und Schüler der Mittelstufe über das Thema Energie unterrichtet wurden, desto komplexer wird ihr Verständnis vom Konzept Energie. Während die Schülerinnen und Schüler in den ersten Jahren der weiterführenden Schule (Jahrgang 5 und 6) zunächst ‚Energieformen und Energiequellen' kennen (Entwicklungsstufe 1), folgen in den darauffolgenden Jahren die Entwicklungsstufe 2 des ‚Energietransfers und der Energieumwandlung', die Stufe 3 der ‚Energieentwertung' und final die Stufe 4 der ‚Energieerhaltung', wenn die Schülerinnen und Schüler die Oberstufe erreichen (Neumann, Viering & Fischer, 2010; Neumann, Viering, Boone & Fischer, 2013).

Basierend auf diesen Entwicklungsstufen entwickelte Viering für die Jahrgänge 6, 8, 10 und 11 Multiple-Choice-Single-Select-Testaufgaben (drei Distraktoren und eine richtige Antwortmöglichkeit), um die einzelnen Entwicklungsstufen abbilden zu können. Dabei wurden neben den inhaltlichen Entwicklungsstufen zusätzlich verschiedene Aufgabenschwierigkeiten konstruiert (nähere Erläuterungen zur inhaltlichen Differenzierung der Entwicklungsstufen und der verschiedenen Aufgabenschwierigkeiten siehe Viering, 2012; Neumann, Viering & Fischer, 2010).

Aus dem Aufgabenpool von Viering (2012) werden für diese Arbeit 22 Aufgaben ausgewählt. Die Auswahl berücksichtigt eine homogene Aufgabenverteilung hinsichtlich der Entwicklungsstufen (alle Stufen sollen erfasst werden) und der Aufgabenschwierigkeit. Da in dieser Studie ein neunter Jahrgang untersucht werden soll, orientiert sich die Aufgabenauswahl zusätzlich an der mittleren Personenfähigkeit eines neunten Jahrgangs. Insgesamt werden auf diese Weise sechs Aufgaben der Entwicklungsstufe 1, fünf Aufgaben der Entwicklungsstufe 2, sechs Aufgaben der Entwicklungsstufe 3 und fünf Aufgaben der Entwicklungsstufe 4 ausgewählt.

Computergestützte Auswertung von Concept Maps

Für eine weitere Validierung der Concept Maps, die von den Lehrpersonen mittels Beurteilungsbogen ausgewertet werden (s. o.), werden die Concept Maps durch ein computerbasiertes Verfahren strukturell ausgewertet. Ziel dieses Verfahrens ist es, die Beurteilung der Concept Maps durch den Beurteilungsbogen, durch das PC-gestützte Verfahren und den Kompetenztest konvergent und diskriminant zu validieren. Zusätzlich kann die Reliabilität der Concept Map-Beurteilung der Lehrpersonen eingeschätzt werden. Die Concept Maps werden für die PC-basierte Auswertung digitalisiert und mit der Computer-Software AKOVIA (Ifenthaler, 2010) nach graphentheoretischen Verfahren ausgewertet. Die Software berechnet strukturelle und semantische Parameter der Concept Maps in Form von Maßzahlen, die über den Vergleich mit einer Durchschnittsmap (Modalmap) erzeugt werden. Für die hier präsentierte Studie werden zwei zentrale semantische Parameter für jede einzelne Concept Map generiert, die Aussagen über die inhaltliche Qualität der Concept Maps geben sollen. Der Parameter Conceptual Matching zählt die Summe der Begriffe, die semantisch der Durchschnittsmap ähnlich sind, der Parameter Propositional Matching die übereinstimmenden Propositionen (vgl. Ifenthaler, 2010). Basierend auf den Definitionen der beiden Parameter ist anzunehmen, dass sie bedingt Aufschluss über die inhaltliche Qualität einer Concept Map geben können. Die Parameter sind mit Einschränkungen für die Interpretation einer Concept Map geeignet.

Tabelle 4.1 zeigt eine Übersicht der eingesetzten Instrumente.

Tabelle 4.1. Eingesetzte Instrumente der Studie 1.

Testinstrument	Quelle
Concept Map-Aufgabenformat A und B	Eigenentwicklung
Concept Map-Beurteilungsbogen, 18 Items	Eigenentwicklung
Kompetenztest (adaptiert)	Viering, 2012
PC-Auswertung (AKOVIA)	Ifenthaler, 2010

4.1.4 Datenerhebung

Die Studie, die in der Zeit zwischen Juni und Juli 2011 stattfand, wurde in zwei Schritten durchgeführt. In deutschen Schulen kann nicht davon ausgegangen werden, dass alle Schülerinnen und Schüler wissen, was Concept Maps sind und wie sie erstellt werden. Um dies abzusichern, wurde zunächst in einer Unterrichtsstunde (45 Minuten) das Concept Mapping eingeübt. In Anlehnung an Sumfleth, Neuroth und Leutner (2010) wurde eine

Stunde konzipiert, in der zunächst gemeinsam mit den Schülerinnen und Schülern ein Prototyp einer Concept Map zum Themengebiet ‚Sehen' erarbeitet wurde. Es folgte eine Übungsphase, in der die Schülerinnen und Schüler eine erste Concept Map zum Themengebiet ‚Magnetismus' selbst erstellten. In der darauffolgenden Reflexionsphase wurde geklärt, ob die Kriterien zur Erstellung einer Concept Map eingehalten wurden und inwiefern die Erstellung den Schülerinnen und Schülern Schwierigkeiten bereitete. Die Erstellung einer weiteren Concept Map zum Thema Aggregatzustände sollte die Vorgehensweise festigen. Den Abschluss der Stunde bildete die Zusammenfassung der Kriterien, auf die bei der Erstellung einer Concept Map geachtet werden soll. Um Lehrereffekte auszuschließen, wurde diese Übungsstunde von der Autorin selbst durchgeführt.

In einem zweiten Termin (90 Minuten), ca. eine Woche nach der Übungsstunde, wurde den Schülerinnen und Schülern zunächst ins Gedächtnis gerufen, was sie aus der letzten Übungsstunde gelernt hatten. Anschließend wurden die Aufgabenformate A und B alternierend verteilt. Schülerinnen und Schüler mit Aufgabenformat A erhielten insgesamt 30 Minuten Zeit für die Bearbeitung. Für Aufgabenformat B hatten die Schülerinnen und Schüler zunächst 15 Minuten Zeit für die Bearbeitung mit den Bildern und anschließend, nach einem Wechsel der Stiftfarbe, weitere 15 Minuten zur Bearbeitung des Aufgabenblatts aus Aufgabenformat A.

Nach 30 Minuten wurde diese Concept Mapping-Phase in der gesamten Klasse beendet und es folgte die Testung der Klasse mit dem Kompetenztest. Alle Schülerinnen und Schüler hatten für die Bearbeitung des Tests 30 Minuten Zeit. Tabelle 4.2 verdeutlicht zusammenfassend das Vorgehen:

Tabelle 4.2. Ablauf der Studie 1.

Vorlauf		**Phase 1**	**Phase 2**	**Phase 3**
Übungsstunde zum Concept Mapping, 45 Minuten	Lerngruppe 1	Aufgabenformat A, 30 Minuten		Kompetenztest, 30 Minuten
	Lerngruppe 2	Aufgabenformat B , Bilder 15 Minuten	Aufgabenformat B, Begriffe, 15 Minuten	

Anschließend wurden die von den Schülerinnen und Schülern erstellten Concept Maps durch die Autorin und durch studentische Mitarbeiter aus dem Hauptstudium des Lehramtsstudiums Physik mit dem Concept Map-Beurteilungsbogen beurteilt. Dabei wurden alle Concept Maps von allen Beurteilern beurteilt, sodass eine Mehrfachkodierung durchgeführt werden konnte.

4.1.5 Ergänzende Schritte nach Studie 1

Nach Durchführung der Studie 1 und einer Betrachtung der Ergebnisse eröffnet sich ein zusätzliches Fragenfeld: Es ist unklar, warum die von den Schülerinnen und Schülern erstellten Concept Maps keine Ergebnisse auf den höheren Kompetenzentwicklungsstufen ‚Energieentwertung' und ‚Energieerhaltung' zeigen. Es ist offen, ob die Schülerinnen und Schüler diese Inhalte nicht kennen, weil sie sie nicht gelernt hatten oder ob sie das Wissen mit der Methode des Concept Mappings nicht ausdrücken können. Um diese Frage zu klären, wird Studie 1 mit zwei Leistungskursen Physik zweier G8-Gymnasien (Jahrgangsstufe 11, Q1) wiederholt. Die Concept Maps der Schülerinnen und Schüler lassen erkennen, dass die Leistungskurse das zu erwartende Verständnis von Energie auf allen vier Kompetenzstufen nach Liu und McKeough (2005) und Neumann, Viering und Fischer (2010) zeigen können. Es kann angenommen werden, dass die Jahrgangsstufe einen Einfluss darauf hat, wie die Concept Maps zum Konzept Energie ausfallen.

Zusätzlich ist nach den mittelmäßigen Ergebnissen zur Beurteilerübereinstimmung der Concept Maps aus Studie 1, die zunächst nur von den studentischen Mitarbeitern bewertet wurden (Ergebnisse siehe in Kapitel 5.1.2) unklar, wie Lehrerinnen und Lehrer Concept Maps beurteilen, wenn sie den Concept Map-Beurteilungsbogen nutzen sollen. Nach den Ergebnissen zur Beurteilerübereinstimmung der studentischen Mitarbeiter kann angenommen werden, dass die studentischen Mitarbeiter mit der Bewertung der Concept Maps fachlich überfordert sind. Deshalb werden zur Bestimmung der Interraterreliabilität alle 79 Concept Maps der Studie 1 zusätzlich von jeweils zwei Lehrerinnen und Lehrern beurteilt, mit dem Auftrag diese mit dem Concept Map-Beurteilungsbogen zu beurteilen. Durch dieses zusätzliche Verfahren kann die Interraterreliabilität allerdings nicht verbessert werden. Nach Wirtz und Caspar (2002) sollten nicht optimale Interraterreliabilitäten, in Abhängigkeit vom untersuchten Merkmal und der Stichprobe betrachtet werden.

4.2 Studie 2

Die Ergebnisse der Studie 1 sind Entscheidungshilfen für das Design und die Auswahl der Instrumente für Studie 2, die in diesem Kapitelabschnitt näher erläutert werden.

4.2.1 Design

Das Ziel der gesamten Studie ist es, festzustellen, wie Lehrkräfte mit den entwickelten Instrumenten unter der Perspektive der Praxistauglichkeit umgehen. Es soll evaluiert werden, wie sich die Nutzung von Concept Maps und Concept Map-Beurteilungsbögen auf die Diagnosegenauigkeit von Physiklehrkräften hinsichtlich einer Diagnose von Schülerkompetenzen im Basiskonzept Energie auswirken. Unter Berücksichtigung der bereits entwickelten Instrumente und der Ergebnisse aus der Vorstudie und der Studie 1 wird ein quasi-experimentelles 2x2-Querschnittsdesign mit den Faktoren ‚Concept Map' und ‚Concept Map-Beurteilungsbogen' gewählt. Das Design ermöglicht die Messung der Diagnosegenauigkeit unter den festgelegten Versuchsbedingungen. Vier verschiedene Gruppen von Schülerinnen und Schülern und ihren Lehrerinnen und Lehrern sollen in verschiedenen Kombinationen die bereits beschriebenen Instrumente in unterschiedlichen Kombinationen nutzen (Abbildung 4.2).

		Mit CM		**Ohne CM**	
Mit CM-BB	LuL	VARIABLE -Rangfolge (Diagnosegenauigkeit), -Kontrollvariablen <u>Gruppe 1</u>	INSTRUMENT -Durch CM-BB -Fragebogen	VARIABLE -Rangfolge (Diagnosegenauigkeit), -Kontrollvariablen <u>Gruppe 3</u>	INSTRUMENT -Durch CM-BB -Fragebogen
	SuS	-Wissensstruktur -Kompetenz, -Intelligenz	-CM -KT -KFT	-Kompetenz, -Intelligenz	-KT -KFT
Ohne CM-BB	LuL	VARIABLE -Rangfolge (Diagnosegenauigkeit), -Kontrollvariablen <u>Gruppe 2</u>	INSTRUMENT -Durch CM -Fragebogen	VARIABLE -Rangfolge (Diagnosegenauigkeit), -Kontrollvariablen <u>Gruppe 4</u>	INSTRUMENT -Durch eigene Maßstäbe -Fragebogen
	SuS	-Wissensstruktur -Kompetenz, -Intelligenz	-CM -KT -KFT	-Kompetenz, -Intelligenz	-KT -KFT

Abbildung 4.2. Studiendesign der Studie 2 zur Messung der Diagnosegenauigkeit von Physiklehrkräften in Abhängigkeit der genutzten Diagnoseinstrumente.
Bemerkungen: CM-BB steht für Concept Map-Beurteilungsbogen, LuL für Lehrerinnen und Lehrer, SuS für Schülerinnen und Schüler, CM für Concept Map, KT für Kompetenztest und KFT für Kognitiver Fähigkeitstest. Lehrkräfte der Gruppen 1 und 2 erhalten anonymisierte Concept Maps zur Rangfolgenbildung. Lehrkräfte der Gruppen 3 und 4 erstellen die Rangfolge mit Hilfe der Namen der Schülerinnen und Schüler, die sie unterrichten.

Alle Lehrpersonen sollen mit den jeweils zur Verfügung stehenden Instrumenten eine Rangfolge der Schülerinnen und Schüler hinsichtlich ihres Verständnisses zum Konzept Energie erstellen. Diese Rangfolge wird im Anschluss mit dem Ergebnis der Schülerinnen und Schüler im Kompetenztest verglichen, das ebenfalls als Rangfolge geordnet werden kann.

In Gruppe 1 erstellen alle Schülerinnen und Schüler anonymisierte Concept Maps zum Konzept Energie, die im Anschluss von der dazugehörigen Physiklehrkraft über den Concept Map-Beurteilungsbogen anonym bewertet werden. Die Lehrkräfte erstellen eine Rangfolge, beginnend mit der Concept Map (=höchste Punktzahl im Beurteilungsbogen), die das breiteste physikalische Verständnis zum Konzept Energie aufzeigt. Pro Untersuchungsgruppe werden nacheinander Beurteilungsbogen und die anonymen Concept Maps als Basis der Beurteilung entfernt. Die vierte Untersuchungsgruppe (Gruppe 4) erhält zur Beurteilung weder die Concept Maps ihrer Schülerinnen und Schüler noch den Concept Map-Beurteilungsbogen. Diese Lehrkräfte bringen ihre eigenen Schülerinnen und

Schüler nach eigenen Kriterien und vorangegangenen Beurteilungen in eine Rangfolge. Dazu benötigen sie die Namen ihrer Schülerinnen und Schüler. Durch das Erstellen der Rangfolgen in allen Versuchsgruppen ist es möglich, die Diagnosegenauigkeit der Lehrkräfte über den Grad an Übereinstimmung zwischen der von den Lehrkräften bestimmten Rangfolge und der Rangfolge der Testleistung zu ermitteln (s. Abbildung 4.2).

An dieser Stelle wird bereits darauf hingewiesen, dass dieses Design kein einwandfreies 2x2-Quasiexperiment ist. Aus Gründen der Durchführung wurden einige Versuchsgruppen nicht untersucht. Weitere Erläuterungen für die Auswahl werden in Kapitel 6 Diskussion für Studie 2 diskutiert.

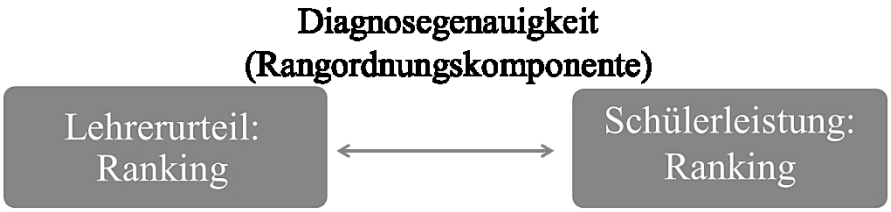


Abbildung 4.3. Zentrale Elemente des Studiendesigns 2.

Als Kontrollvariablen werden zusätzliche Schüler- und Lehrermerkmale berücksichtigt. Schülermerkmale sind beispielsweise die kognitive Fähigkeit oder Physiknote. Auf Lehrerebene können die Anzahl der Berufsjahre und das Alter einen Einfluss auf die Diagnosegenauigkeit haben. Tabelle 4.3 fasst die relevanten Kontrollvariablen zusammen.

Tabelle 4.3. Relevante Kontrollvariablen auf Schüler- und Lehrerebene.

Ebene	Kontrollvariable	Instrument
Schülerinnen und Schüler	Intelligenz	KFT
	Schulnoten	Kompetenztest
Lehrerinnen und Lehrer	Anzahl der Berufsjahre	Lehrerfragebogen
	Alter	
	Geschlecht	
	Kenntnis von Concept Maps	
	Nutzung von Concept Maps	

4.2.2 Stichprobe

Da die ergänzenden Schritte nach Studie 1 gezeigt haben, dass eine inhaltliche Qualitätssteigerung der Concept Maps mit zunehmendem Jahrgang erwartet werden kann,

wird in dieser Studie der Fokus auf die Einführungsphase (EF) der reformierten G8 Gymnasien gelegt (jetzt Klasse 10, EF). Es kann davon ausgegangen werden, dass Schülerinnen und Schüler der Einführungsphase nach dem neunten Jahrgang weiteres Verständnis zum Konzept Energie entwickelt haben. Da dieses Projekt nordrhein-westfälische Gymnasien in ihrer Entwicklung zum Ganztagsgymnasium begleitet, wird als Schulform das Gymnasium ausgewählt, um die Projektschulen als Teilnehmer an der Studie zu gewinnen. Darüber hinaus wird diese Studie weiteren Gymnasien angeboten.

Die Stichprobe besteht aus 48 Physiklehrerinnen und Physiklehrern mit 977 Schülerinnen und Schülern aus 38 Schulen. Die Erhebung wurde in zwei Schulhalbjahren durchgeführt. Sie begann mit dem zweiten Halbjahr des Schuljahres 2011/12 und wurde nach den Sommerferien im ersten Schulhalbjahr 2012/13 weitergeführt. Annähernd die gesamte Stichprobe der Gruppe 3 wurde im ersten Schulhalbjahr 2012/13 getestet. Diese Lehrkräfte kannten ihre Kurse im Verhältnis zu den Lehrkräften, die im zweiten Halbjahr des Schuljahres 2011/12 teilgenommen haben, kurzzeitig. Die Mehrheit der Lehrkräfte wurde im Halbjahr vor den Sommerferien getestet.

4.2.3 Beschreibung der Instrumente

Für diese Studie wird auf das bereits beschriebene papierbasierte *Concept Map-Aufgabenformat*, den *Kompetenztest* und den *Concept Map-Beurteilungsbogen* zurückgegriffen.

Die Ergebnisse der Studie 1 zeigen, dass hinsichtlich des Ziels einer zeitökonomischen Nutzung des Bogens eine Kürzung des Concept Map-Beurteilungsbogens auf weniger als 18 Aussagen erforderlich ist (vgl. Kapitel 5 Ergebnisse zur Zeitökonomie in den Studien 1 und 2). Mit einer Faktorenanalyse wurden inhaltlich gleiche Aussagen im Bogen ermittelt und die redundanten Aussagen entfernt, sodass der Beurteilungsbogen für diese Studie nur noch aus 10 Aussagen besteht. Dadurch können Lehrkräfte zeitökonomischer arbeiten. Aufgabenformat B wird eingesetzt, da es inhaltlich umfangreichere Concept Maps für Diagnosezwecke erzeugen kann als das Aufgabenformat A (vgl. hierzu das Kapitel 5.1.2 Analyseergebnisse-Gruppenunterschiede im Concept Mapping). Zusätzlich werden die in Tabelle 4.3 genannten Kontrollvariablen erhoben. Die Lehrpersonen der Gruppen 1 und 3 erhalten darüber hinaus Instruktionen zur Nutzung des (Concept Map-) Beurteilungsbogen (siehe als Zusammenfassung Tabelle 4.4).

Tabelle 4.4. Eingesetzte Instrumente der Studie 2.

Testinstrument	Quelle
Concept Map-Aufgabenformat A	Eigenentwicklung
Concept Map-Beurteilungsbogen, 10 Items	Eigenentwicklung
Kompetenztest	Viering, 2012
Kognitiver Fähigkeitstest	Heller & Perleth, 2000
Lehrerfragebogen zu Ausbildung und Beruf	Eigenentwicklung
Lehrerinnen-und Lehrer-Rankingbogen	Eigenentwicklung
Manual zur Nutzung des Concept Map-Beurteilungsbogens der Gruppen 1 und 3	Eigenentwicklung

Kognitiver Fähigkeitstest, KFT

Durch den Kognitiven Fähigkeitstest-Revision (kurz: KFT-R) nach Heller und Perleth (2000) wird der Einfluss kognitiver Fähigkeiten der Schülerinnen und Schüler auf die gezeigte Leistung in den Concept Maps und dem Kompetenztest kontrolliert. Es wird von dem Gesamttest, der sich in drei Skalen gliedert (verbale Fähigkeiten, quantitative Fähigkeiten und nonverbale Fähigkeiten), die nonverbale Unterskala N2, Form A für die Jahrgangsstufe 10 eingesetzt. In diesem Test soll die Fähigkeit des figuralen Denkens ermittelt werden, die mit der Fähigkeit Concept Maps zu erstellen, zusammenhängen kann. Auf die Form A wird zurückgegriffen, da nach einer Untersuchung von Segerer, Marx und Marx (2012) die Skala N2 der Form B zwei unlösbare Aufgaben beinhaltet. Die eingesetzte Unterskala soll in 8 Minuten bearbeitet werden (vgl. Heller & Perleth, 2000).

Lehrerfragebogen zu Ausbildung & Beruf

Um mögliche Einflüsse auf die Diagnosegenauigkeit einer Lehrkraft festzustellen, werden Lehrermerkmale in Form von Kontrollvariablen in dieser Studie erfasst. Hierzu werden die Lehrerinnen und Lehrer in einem selbstentwickelten Lehrerfragebogen nach ihrem demografischen Hintergrund, etwa dem Studienabschluss und der Anzahl der Berufsjahre befragt. Tabelle 4.3 im Abschnitt Design fasst die wesentlichen Kontrollvariablen auf Lehrerebene zusammen.

Lehrerinnen- und Lehrer-Rankingbogen

Je nach Gruppenzugehörigkeit stehen den Lehrerinnen und Lehrern verschiedene Instrumente zur Beurteilung ihrer Schülerinnen und Schüler zur Verfügung (vgl. Abbildung 4.2, z. B. Concept Maps ‚ja oder nein'). Die von den Lehrerinnen und Lehrern generierte Rangfolge über die Schülerinnen und Schüler basiert daher auf verschiedenen

Grundlagen. Alle Lehrpersonen notieren auf einem Rankingbogen die von ihnen ermittelte Rangfolge. Die Lehrerinnen und Lehrer erhalten im Rankingbogen kurze Instruktionen, wie sie die Rangfolge für ihre Gruppe jeweils genau zu erstellen haben. Die Rankingbögen der einzelnen Gruppen werden im Anhang unter A.5 erläutert.

Manual zur Nutzung des (Concept Map)-Beurteilungsbogens der Gruppen 1 und 3

Die Nutzung des (Concept Map)-Beurteilungsbogens ist nur teilweise selbsterklärend. Es wird für die Lehrerinnen und Lehrer der Gruppe 1 ein Manual entwickelt, in dem an Beispiel-Concept Maps und kurzen Beschreibungen erklärt wird, wie der Beurteilungsbogen genutzt werden soll.

Die Lehrerinnen und Lehrer der Gruppe 3 (keine Concept Maps, aber Beurteilungsbogen) erhalten ebenfalls das Manual für die Nutzung des Beurteilungsbogens. Dieses Manual beinhaltet leicht abgewandelte Instruktionen ohne Beispiel-Concept Maps, da den Lehrkräften dieser Gruppe keine Concept Maps zur Verfügung stehen. Den Lehrerinnen und Lehrern wird erklärt, dass sie den Bogen auf Basis ihrer Erfahrung mit der jeweiligen Schülerin/ dem jeweiligen Schüler ausfüllen sollen. Die Manuale werden im Anhang unter A.4 beschrieben.

4.2.4 Datenerhebung

Die Studie wurde im Zeitraum zwischen März 2012 und Januar 2013 durchgeführt. Insgesamt stellten die vier verschiedenen Gruppen unterschiedliche Anforderungen an die Umsetzung.

Der nachfolgende Ablauf (Tabelle 4.5) gibt einen Überblick über das Vorgehen in den einzelnen Gruppen *während der Schulbesuche.*

Tabelle 4.5. Ablauf der Studie 2.

Besuch A		**Besuch B**			
Vorlauf		**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**
Übungsstunde zum Concept Mapping, 45 Minuten	Gruppe 1	Aufgabenformat B, Bilder, 15 Minuten	Aufgabenformat B, Begriffe, 15 Minuten	Kompetenztest, 30 Minuten	KFT, 8 Minuten
	Gruppe 2	Aufgabenformat B, Bilder, 15 Minuten	Aufgabenformat B, Begriffe, 15 Minuten		
-	Gruppe 3	-	-		
-	Gruppe 4	-	-		

Bemerkungen: Allen Gruppen ist Phase 3 und 4 gemeinsam.

Die in Studie 1 bereits erprobte Übungsstunde zum Concept Mapping wurde in einem ersten Besuchstermin (Besuch A) in den Gruppen 1 und 2 durchgeführt. Im Anschluss folgte in einem zweiten Besuchstermin (Besuch B) die Erhebung, in der den Schülerinnen und Schülern beider Gruppen das Aufgabenformat B, gefolgt vom Kompetenztest und dem KFT, zur Bearbeitung gegeben wurde. Für diesen zweiten Besuchstermin wurden 90 Minuten benötigt. Die Teilnehmer der Gruppen 3 und 4 wurden jeweils einmal besucht. Die Schülerinnen und Schüler dieser Gruppen bearbeiteten in insgesamt 45 Minuten den Kompetenztest und den KFT (vgl. Tabelle 4.5).

Vor der eigentlichen Erhebung erhielten alle Schülerinnen und Schüler ein mit einer Nummer bedrucktes Kärtchen, auf deren Rückseite sie ihren Namen schreiben sollten. Diese individuellen Nummern wurden von den Schülerinnen und Schülern auf allen ausgeteilten Materialien notiert. Da die Lehrpersonen der Gruppen 3 und 4 keine Concept Maps zur Bildung der Rangfolge hatten, mussten ihnen die Kärtchen mit den Namen nach den Erhebungen zur Verfügung stehen (vgl. Abschnitt 4.2.1 Design). Sie hatten keine weitere Information für die Bildung einer Rangfolge und mussten sich auf ihre Erfahrungen mit den Schülerinnen und Schülern aus vorangegangenem Unterricht stützen, zu der sie die Namen benötigten. Die Lehrkräfte der Gruppen 1 und 2 benötigten diese Namenskärtchen nicht. Um möglichst viele Faktoren in der Durchführung konstant zu halten, wurden den Lehrkräften ebenfalls die Kärtchen überlassen. Es wurde diesen Lehrpersonen jedoch gesagt, dass die Kärtchen für sie keine Bedeutung in der Bewertung der Concept Maps haben. Zusätzlich wurden die Concept Maps durch dieses Verfahren anonymisiert. Die Kärtchen wurden nach der Untersuchung von allen Lehrkräften vernichtet.

Nach der Datenerhebung in den Schulen erhielten alle teilnehmenden Lehrkräfte per Post Anweisungen für die Bildung der Rangreihen ihrer Schülerinnen und Schüler. Die Lehrkräfte aus den Gruppen 1 und 2 erhielten die von ihren Schülerinnen und Schülern erstellten anonymen Concept Maps. Gruppe 1 bekam zusätzlich die Concept Map-Beurteilungsbögen mit dem Manual. Gruppe 2 erhielt bis auf die Concept Maps keine weiteren Hilfestellungen. Gruppe 3 bekam die Beurteilungsbögen mit dem Manual und Gruppe 4 keine Hilfestellungen (vgl. Abb. 4.2). Alle vier Gruppen erhielten den Rankingbogen, auf dem sie ihre Rangfolge notieren konnten, und den Lehrerfragebogen.

Zur Durchführung der Erhebung wurden nach einer Testleiterschulung studentische Mitarbeiter eingesetzt. Die Autorin übernahm die Concept Map-Übungsstunden der Gruppen 1 und 2, während die studentischen Mitarbeiter die übrigen Termine wahrnahmen.

4.3 Statistische Methoden und Datenanalyse

Merkmale von Personen können über Fragebögen und Tests gemessen werden. Die Entwicklung von Tests und die Auswertung dieser Daten kann über zwei grundsätzliche Testtheorien erfolgen: die klassische Testtheorie und die probabilistische Testtheorie (Bühner, 2006). Eine Entscheidung für eine Testtheorie hängt davon ab, was aus dem späteren Datensatz erfahren werden soll und ob die jeweilige Testtheorie sinnvolle Interpretationen erlaubt (vgl. u. a. Darstellung der Unterschiede in Bühner, 2006; Bortz & Döring, 2006).

In dieser Arbeit erfolgt die Auswertung der Daten nach den Methoden der klassischen Testtheorie. Durch sie können Rangdaten, wie sie in dieser Arbeit vorliegen (in Studie 2), ausgewertet werden. Die probabilistische Testtheorie hingegen setzt für eine Analyse die Beantwortung von Testitems voraus, um durch die Analyse von Antwortmustern auf die latente Fähigkeit einer Person schließen zu können (vgl. Bühner, 2006). Die probabilistische Testtheorie ist für die Auswertung der in dieser Arbeit vorliegenden Daten (Rangdaten) ungeeignet.

Die statistischen Analysen, die zur Beantwortung der beiden Forschungsfragen benötigt werden, werden durch *deskriptive Statistiken*, beispielsweise der Beschreibung der Teilnehmerstruktur in ihrer Geschlechterzusammensetzung oder des Alters ergänzt. Es wird ein Überblick ermöglicht, der Hilfe für eine Interpretation der Ergebnisse sein kann.

4.3.1 Studie 1

Um statistische Analysen durchführen zu können, müssen die Daten bestimmte Voraussetzungen erfüllen. Ein Kriterium ist die *Normalverteilung* des erhobenen Merkmals. In dieser Studie wird die Leistung der Schülerstichprobe im Kompetenztest graphisch und durch den Kolmogorov-Smirnov-Test (*K*-*S*-Test) auf Normalverteilung untersucht. Es kann von einer Normalverteilung ausgegangen werden, wenn der *K*-*S*-Test nicht signifikant wird. Folgt die Schülerleistung keiner Normalverteilung, müssen die statistischen Tests, die mit der Schülerleistung in Verbindung stehen, mit verteilungsfreien, sogenannten nicht-parametrischen Verfahren durchgeführt werden.

Ein weiteres Kriterium für statistische Tests ist die *Varianzhomogenität*. Mit dem Levene-Test wird die Gleichheit der Varianzen in den Schülerstichproben geprüft, die die unterschiedlichen Concept Map-Aufgabenformate bearbeiten. Ein signifikantes

Testergebnis deutet darauf hin, dass keine Gleichheit der Varianzen in den Gruppen angenommen werden kann. Wird dieses Kriterium nicht erfüllt, müssen die weiteren Tests ebenfalls nicht-parametrisch durchgeführt werden.

Das Gütekriterium *Reliabilität des Concept Map-Beurteilungsbogens* wird in Form einer Interraterübereinstimmung ermittelt. Bei diesem Verfahren soll ermittelt werden, wie groß der Fehler ist, der durch die Anwendung des Beurteilungsbogens entsteht (Reliabilität). Er vergleicht verschiedene Beurteiler, die die gleichen Concept Maps beurteilen. Da der Beurteilungsbogen intervallskalierte Daten produziert, wird als Übereinstimmungsmaß der justierte Interklassen-Korrelations-Koeffizient, zwei-Wege-gemischt-Modell für randomisierte Beurteiler gewählt (ICC_{just}). Der ICC_{just} ermöglicht es, ein Gesamtmaß für alle sechs Rater berechnen zu können, die alle 79 Concept Maps beurteilen (vgl. Wirtz & Caspar, 2002). Der ICC_{just} kann Werte zwischen Null und 1 annehmen. Ist der ICC_{just} gleich Null, besteht keine Übereinstimmung zwischen den Urteilen mehrerer Beurteiler. Nähert sich der Wert 1, kann davon ausgegangen werden, dass die Beurteilungen zunehmend reliabel sind (u. a. Wirtz & Caspar, 2002; Weir, 2005). Wirtz und Caspar (2002) weisen darauf hin, dass „allgemein […] in der Literatur eine Interklassenkorrelation von mindestens .7 als Indiz für ‚gute' Reliabilität angesehen [*wird*] (Greve & Wentura, 1995). Dies kann jedoch nur eine sehr vage Richtlinie sein, da […] die Ausprägung der Koeffizienten immer in Abhängigkeit vom zu messenden Merkmal und der untersuchten Stichprobe beurteilt werden muss. […]" (Wirtz & Caspar, 2002, 160).

Das Übereinstimmungsmaß Cohens κ ist für diese Studie ungeeignet, da es keine Gesamtübereinstimmung zwischen mehr als zwei Ratern ermitteln kann. Das ordinale Übereinstimmungsmaß γ eignet sich nicht, da es als ein sehr mildes Maß eingeschätzt wird und die Reliabilität zwischen Ratern positiv verzerrt abbildet.

Neben der Objektivität des Beurteilungsbogens wird mit der *internen Konsistenz (Homogenität)* geprüft, inwiefern seine verschiedenen Aussagen dasselbe Konstrukt messen-die Erfassung der Kompetenz im Basiskonzept Energie. Der Kompetenztest wird ebenfalls auf interne Konsistenz untersucht. Das zu wählende Maß für beide Fälle ist Cronbachs α, dessen Werte zwischen -1 und 1 liegen können. Werte, die größer als .7 sind, können als akzeptabel eingestuft werden (vgl. Rost, 2005; Field, 2009).

Um die *konvergente Validität* zwischen Concept Maps und dem Kompetenztest ermitteln zu können, werden diese beiden Instrumente durch eine Korrelation verglichen. Ist die Schülerleistung normalverteilt, wird die Korrelation über Pearsons Korrelationskoeffizienten *r* für intervallskalierte Daten berechnet. Das Analogon für eine

nicht normalverteilte Schülerleistung ist das nicht-parametrische Verfahren mit Spearmans Rangkorrelationskoeffizienten ρ. Im Gegensatz zu Pearsons Korrelationskoeffizient r werden zur Berechnung von Spearmans ρ nicht die tatsächlich gemessenen Schülerleistungsdaten als Berechnungsbasis genommen, sondern die Schülerleistung wird in Ränge transformiert. Beide Korrelationskoeffizienten können Werte zwischen -1 und 1 annehmen, wobei bei einem Wert von 1 von einer perfekten Korrelation gesprochen wird. In diesem Fall würden beide zu testenden Instrumente das gleiche Konstrukt, die Kompetenz im Themengebiet ‚Energie', messen. Zusätzlich wird über eine weitere Korrelationsberechnung die PC-gestützte Auswertung mit dem Kompetenztest und dem Concept Map-Beurteilungsbogen zur Triangulation verglichen.

In einer weiteren Analyse soll überprüft werden, ob sich die zwei Schülergruppen unterscheiden, die die zwei unterschiedlichen Concept Map-Aufgabenformate bearbeiten. Wird von einer normalverteilten Schülerleistung im Kompetenztest ausgegangen, wird aus den individuellen Beurteilungen aus den Beurteilungsbögen für jede Schülergruppe der Punktemittelwert ermittelt. Die zwei Gruppenmittelwerte werden über den *t*-Test für unabhängige Stichproben miteinander verglichen. Die Ergebnisse des *t*-Tests können Aussagen über die Aufgabenformate und deren Potenziale für eine Schülerdiagnose liefern. Derselbe statistische Test lässt sich für einen Mittelwertvergleich dieser beiden Schülergruppen hinsichtlich der erreichten Punkte im Kompetenztest anwenden. In beiden Fällen kann der *relative Effekt* im *t*-Test über den Effektstärke-Test nach Cohen (1988) verdeutlicht werden. Es wird davon ausgegangen, dass das berechnete Effektstärkemaß Cohens *d* mit $d > .8$ einen großen Effekt kennzeichnet, während bei Werten von $d > .5$ von einem mittleren Effekt gesprochen wird und bei $d < .2$ von einem kleinen Effekt (Cohen, 1988; Bühner, 2006).

Das Pendant zum *t*-Test, das auf eine spezielle Verteilungsannahme für die Grundgesamtheit der Stichprobe verzichtet, ist der Mann-Whitney-*U*-Test (*U*-Test). Im Falle einer nicht normalverteilten Schülerleistung werden Rangplätze, die die Schülerinnen und Schüler auf Grund ihrer Schülerleistung erhalten, als Vergleichsbasis gewählt. Das Ergebnis des *U*-Tests kann wie das Ergebnis des *t*-Tests interpretiert werden. Die Effektstärke für verteilungsfreie Verfahren wird durch ω dargestellt. Nach Bühner und Ziegler (2009) weisen ω-Werte bis .2 kleine Effekte, ω bis .4 moderate und ω ab .5 große Effekte auf.

Um die in den statistischen Tests gefundenen Effekte statistisch optimal abzusichern, werden nach der Durchführung der primär interessierenden Tests post-hoc-

Teststärkeberechnungen durchgeführt. Die Teststärke 1-β gibt an, mit welcher Wahrscheinlichkeit der eingesetzte Test den angenommenen Effekt gefunden hat, falls dieser tatsächlich existiert. Das β gibt an, mit welcher Wahrscheinlichkeit der eingesetzte Test den tatsächlich vorhandenen Effekt nicht aufdeckt und übersieht. Dadurch können zusätzliche Aussagen getroffen werden, inwiefern eine Stichprobenvergrößerung die Wahrscheinlichkeit erhöht, den existierenden Effekt wirklich zu finden. Tabelle 4.6 fasst die verwendeten statistischen Tests zusammen.

Tabelle 4.6. Zusammenfassung der genutzten statistischen Tests der Studie 1.

Ziel	Statistischer Test		Basis
Normalverteilung der Stichprobe	Kolmogorov-Smirnov-Test und graphische Auswertung		Schülerleistung der Gesamtstichprobe im Kompetenztest
Varianzhomogenität der Gruppen	Levene-Test		Schülerleistung der zwei Gruppenstichproben im Kompetenztest
Objektivität/ Interraterreliabilität	Intraklassen-Korrelationskoeffizient ICC_{just}		Sechsfach-Rating von 79 Concept Maps
Interne Konsistenz	Cronbachs α		a) 474 Concept Map-Beurteilungsbögen von 79 Concept Maps
			b) Kompetenztest der Gesamtschülerstichprobe
Konvergente Validität	parametrisch: Pearsons Produkt-Moment-Korrelationskoeffzient r	nicht-parametrisch: Spearmans Rangkorrelations-koeffizient ρ	a) Leistung im Kompetenztest und Concept Map-Beurteilungsbogen
			b) PC-Auswertung (AKOVIA), Leistung im Concept Map-Beurteilungsbogen und Kompetenztest
Gruppenunterschied	parametrisch: t-Test für unabhängige Stichproben	nicht-parametrisch: Mann-Whitney-U-Test	a) Leistung im Concept Map-Beurteilungsbogen für die Schülergruppen mit Aufgabenformat A und B
			b) Leistung im Kompetenztest für die Schülergruppen mit Aufgabenformat A und B
Größe eines Effekts	parametrisch: Cohens d	nicht-parametrisch: ω	Gruppenunterschied
Teststärke	$1-\beta$		Größe des Effektes

4.3.2 Studie 2

Wie in Studie 1 wird für den Kompetenztest ebenfalls die *interne Konsistenz* in Form von Cronbachs α ermittelt. Zusätzlich wird sie für den eingesetzten kognitiven Fähigkeitstest (KFT) für alle teilnehmenden Schülerinnen und Schüler geprüft. Die Berechnungen zur internen Konsistenz werden für den Concept Map-Beurteilungsbogen ergänzt, der in den Gruppen 1 und 3 genutzt wird. Es werden ebenfalls die Testvoraussetzungen in Form einer *Normalverteilungsprüfung* für die Schülerstichprobe im Kompetenztest und KFT überprüft. Hierzu werden erneut der Kolmogorov-Smirnov-Test und eine graphische Betrachtung vorgenommen.

Die Herausforderung dieser Studie ist es, ein Maß zu erhalten, das Aussagen über die *Diagnosegenauigkeit* einer Lehrkraft ermöglicht. Als ein Weg, dieses Maß zu erhalten, gelten Korrelationen (vgl. Schrader, 1989). Hierbei wird die Rangfolge, die eine Lehrkraft über seine Lerngruppe erstellt, mit der Rangfolge der Lerngruppe verglichen, die sie auf Basis ihrer erbrachten Leistung erhält. In dieser Studie werden die Rangreihen über Spearmans ρ korreliert. Spearmans Test wird gewählt, da die Rangfolgedaten in Form einer ordinalen Skala vorliegen. Jede Lehrkraft erhält über diese Rangkorrelation einen Korrelationswert, der als Maß für die Güte der Diagnosegenauigkeit angesehen werden kann. Dieses Maß ist intervallskaliert. Mit dem Kolmogorov-Smirnov-Test werden die ermittelten Korrelationsmaße (=Diagnosegenauigkeitsmaße) aller Lehrerinnen und Lehrer auf *Normalverteilung* untersucht. Außerdem wird der Levene-Test eingesetzt, um die *Varianzhomogenität* in den Lehrergruppen zu überprüfen.

Wie sich die vier verschiedenen Gruppen hinsichtlich ihrer *Diagnosegenauigkeit* statistisch unterscheiden lassen, lässt sich bei normalverteilten Daten zur Diagnosegenauigkeit mit einer einfaktoriellen Varianzanalyse (ANOVA oder *F*-Test genannt) untersuchen. Die Basis der Analyse stellen die zuvor ermittelten Rangkorrelationsmaße. Die einfaktorielle Varianzanalyse ermittelt einen Mittelwert der Korrelationswerte aller sich in einer Gruppe befindlichen Lehrkräfte und vergleicht diese miteinander. Für die Einschätzung eines *Gruppenunterschieds* wird als Effektstärkemaß eta^2 berechnet. eta^2 kann zwischen Null und 1 rangieren. Die Konvention nach Cohen (1988) bemisst eta^2-Werte bis .01 mit einem kleinen Effekt, eta^2-Werte von .06 mit einem mittleren Effekt und Werte größer gleich .14 deuten auf einen großen Effekt hin (vgl. Sedlmeier & Renkewitz, 2008). Die *einzelnen Gruppen* werden post hoc in ihrer Diagnosegenauigkeit verglichen. Cohens *d* (1988) ermöglicht hierbei eine Einschätzung

der Effektstärke des Gruppenunterschieds zwischen einzelnen Paaren (vgl. *t*-Test in Studie 1 zwischen den zwei Schülergruppen).

Im Falle von nicht-normalverteilten Rangkorrelationen wird statt des *F*-Tests die Rangvarianzanalyse nach Kruskal und Wallis (*H*-Test) eingesetzt. Verteilungsfreie Verfahren, wie der *H*-Test, transformieren die gemessenen Merkmalsausprägungen, hier die Rangkorrelationen (=Diagnosegenauigkeiten), in Ränge und vergleichen die verschiedenen Lehrergruppen auf Basis dieser Ränge. Die Effektstärke des Gruppenunterschieds bezogen auf die Diagnosegenauigkeit wird über ω angegeben, das wie das ω beim *U*-Test zwischen zwei Gruppen interpretiert werden kann (vgl. Bühner & Ziegler, 2009; Field, 2009). Post hoc-Einzelgruppenvergleiche werden wie der Schülervergleich in Studie 1, über *U*-Tests berechnet. Das kritische Signifikanzlevel $p = .05$ wird bei vielen einzelnen Gruppenvergleichen nach der Bonferroni-Korrektur auf $p_{korrigiert} = .05/6$ adjustiert (sechs steht für sechs interessierende Vergleiche). Durch dieses Vorgehen wird eine Inflation des kritischen Signifikanzlevels unterbunden (vgl. Field, 2009).

Zusätzlich wird die Teststärke 1-β der parametrischen und nicht-parametrischen Tests ermittelt. Dadurch können die Ergebnisse der Tests bezüglich ihrer Relevanz interpretiert werden (vgl. Beschreibung zur Teststärke im Abschnitt 4.3.1. Studie 1).

Die zusätzlich erhobenen *Kontrollvariablen*, wie das Alter der Lehrkräfte oder die Berufserfahrung können einen Einfluss auf die Höhe der Diagnosegenauigkeit der Lehrerinnen und Lehrer haben. Um ihren zusätzlichen Effekt auf die Diagnosegenauigkeit zu vermeiden, werden sie durch eine Kovarianzanalyse (auch ANCOVA) kontrolliert bzw. ‚neutralisiert'. Eine parametrische ANCOVA setzt normalverteilte, intervall- oder nominal-skalierte Daten voraus, die unabhängig von dem Gruppeneffekt sind (vgl. Field, 2009). Dies wird durch Korrelationsberechnungen nach Pearson geprüft. Ist die Unabhängigkeit gewährleistet, kann die ANCOVA durchgeführt werden. Ihr möglicher Effekt auf die Diagnosegenauigkeit wird kontrolliert und herausgerechnet, indem die zuvor beschriebene einfaktorielle ANOVA um die weiteren Kontrollvariablen ergänzt wird. Die anschließend ermittelten Ergebnisse können wie bei der einfaktoriellen ANOVA interpretiert (vgl. Sedlmeier & Renkewitz, 2008) und damit Aussagen über den Einfluss der Gruppenzugehörigkeit gemacht werden.

Liegen nicht-normal verteilte Daten vor, werden Korrelationen nach Spearman berechnet, um einen Zusammenhang zwischen den Kontrollvariablen und der

Diagnosegenauigkeit der Lehrkräfte herzustellen. Sie ermöglichen Aussagen über mögliche Zusammenhänge auf einer allgemeinen Ebene.

Für die Schülerstichprobe wird angenommen, dass die Kontrollvariablen einen Zusammenhang mit der Schülerleistung im Kompetenztest aufweisen. Ist die Schülerleistung im Kompetenztest normalverteilt, wird zur Variablenkontrolle in dieser Situation nicht wie bei den Lehrkräften die Kovarianzanalyse genutzt, sondern die Regressionsanalyse. Die Regressionsanalyse erlaubt Aussagen, inwiefern die Kontrollvariablen die Schülerleistung im Kompetenztest erklären können. Dadurch wird es möglich, den Effekt der Kontrollvariablen einzuschätzen. Vor der Regressionsanalyse werden mögliche Zusammenhänge zwischen den Variablen durch eine Korrelation untersucht. Die anschließende eigentliche Regressionsanalyse baut stufenweise Kontrollvariablen in das Analysemodell ein. Mit diesem Vorgehen kann sukzessive der Anteil der Kontrollvariablen auf die Leistung im Kompetenztest erklärt werden. Kontrollvariablen, die einen Einfluss auf die Schülerleistung im Kompetenztest haben können, sind beispielsweise die kognitive Fähigkeit und die letzten Schulnoten in Physik, Deutsch und Mathematik. Diese müssen intervallskaliert sein. Sind die Schülerleistungsdaten nicht normalverteilt, wird der Zusammenhang der Kontrollvariablen mit der Schülerleistung im Kompetenztest über den Korrelationskoeffizienten nach Spearman dargestellt.

Ein möglicher Zusammenhang der KFT-Leistung mit der Leistung in den Concept Maps wird bei den Schülerinnen und Schülern der Gruppe 1 über eine Korrelation ermittelt. Die erreichten Punkte im KFT werden mit den Punkten, die den Concept Maps im Beurteilungsbogen gegeben werden, nach Pearson (falls eine normalverteilte Schülerleistung im KFT vorliegt) bzw. nach Spearman (für eine nicht-normalverteilte Schülerleistung im KFT) korreliert. Ein hoher signifikanter Korrelationswert weist auf einen engen Zusammenhang der Variablen hin. Die KFT-Leistung wäre in diesem Fall von der Leistung in den Concept Maps nicht vollends trennbar.

Experimentelle Designs, speziell mehrfaktorielle Designs, können eine Wechselwirkung zwischen den verschiedenen Faktoren sichtbar machen. Auf Basis der Datenstruktur in diesem Projekt ist es zusätzlich möglich, *Haupteffekte* und *Interaktionen* der Faktoren *Concept Map-Aufgabenformat* und *Concept Map-Beurteilungsbogen* bezogen auf die mittlere Diagnosegenauigkeit einer jeden Versuchsgruppe zu ermitteln. Hierbei wird jeder Lehrperson in Abhängigkeit der Gruppenzugehörigkeit die Ausprägung des jeweiligen Faktors zugeordnet, d. h. Lehrerinnen und Lehrer der Gruppe 1: Concept Map-

Aufgabenformat- ja/ Concept Map-Bewertungsbogen- ja, Lehrpersonen der Gruppe 3 haben als Faktorausprägungen: Concept Map-Aufgabenformat- nein/ Concept Map-Bewertungsbogen- ja usw. Diese Neugruppierung ermöglicht Aussagen zu treffen, inwiefern die Wirkung des einen Faktors von der Ausprägung des anderen Faktors abhängig ist. Die Zusammenhänge können in einer zweifaktoriellen ANOVA und graphisch in Form von Profilplots ermittelt werden. Die dargestellte Berechnung der Haupteffekte und Interaktionen setzt eine Normalverteilung der Diagnosegenauigkeit der Lehrerinnen und Lehrer voraus. Tabelle 4.7 stellt eine Zusammenfassung der genutzten Testverfahren der Studie 2 dar.

Parametrische Verfahren sind robust gegen die Verletzung von Testvoraussetzungen wie z. B. einer nicht vorhandenen Normalverteilung (vgl. Bühner & Ziegler, 2009). Die Entscheidung für die Verwendung von parametrischen und nicht-parametrischen Verfahren wird in dieser Arbeit an den entsprechenden Stellen getroffen. Es wird jeweils das angemessene Testverfahren genutzt und diskutiert. Die Teststärkeberechnungen werden mit der Software G*Power 3.1.7 durchgeführt (Faul, Erdfelder, Lang & Buchner, 2007). Die Effektstärkeberechnungen Cohens *d* über die Homepage http://ncalculators.com/statistics/effect-of-size-calculator.htm (letzter Zugriff am 09.10.2013). Alle weiteren Berechnungen werden mit der Statistiksoftware IBM SPSS Statistics Version 18 und Version 21 durchgeführt (IBM 2012, 2010; vgl. Bühl, 2010). In der Software SPSS (beide Versionen) können keine Teststärken und die Effektstärke nach Cohen berechnet werden. Aus diesem Grund werden diese statistischen Tests mit den vorgenannten Programmen ermittelt.

Tabelle 4.7. Zusammenfassung der genutzten statistischen Tests der Studie 2.

Ziel	Statistischer Test		Basis
Interne Konsistenz	Cronbachs α		a) Kompetenztest der Gesamtschülerstichprobe
			b) KFT der Gesamtschülerstichprobe
			c) Concept Map-Beurteilungs-bögen der Gruppe 1
			d) Concept Map-Beurteilungs-bögen der Gruppe 3
Korrelationsmaß für Diagnosegenauigkeit	Spearmans ρ		Rangfolge der Lehrkraft und Rangfolge durch den Kompetenztest
Normalverteilung der Stichproben	Kolmogorov-Smirnov-Test und graphische Auswertung		a) Schülerleistung der Gesamtstichprobe im Kompetenztest
			b) KFT der Gesamtstichprobe
			c) Diagnosegenauigkeits-leistung aller Lehrkräfte
Varianzhomogenität der Gruppen	Levene-Test		Diagnosegenauigkeit der Lehrerkräfte in den einzelnen Gruppen
Gruppenunterschied	parametrisch: ANOVA (*F*-Test und Post-hoc LSD)	nicht-parametrisch: Rangvarianzanalyse (*H*-Test und *U*-Test)	Diagnosegenauigkeit der Lehrkräfte in den einzelnen Gruppen
Größe eines Effekts	parametrisch: eta^2 und d	nicht-parametrisch: ω	Gruppenunterschied
Teststärke	$1-\beta$		Größe des Effektes
Einfluss der Kontrollvariablen	parametrisch: ANCOVA, Regressionsanalyse und Pearsons Produkt-Moment-Korrelationskoeffizient r	nicht-parametrisch: Spearmans Rangkorrelationskoeffizient ρ	a) Lehrerfragebogen, Diagnosegenauigkeit der Lehrkräfte
			b) Kompetenztest, Schulnoten, Alter
			c) KFT, Punkte aus Concept Map-Beurteilungsbogen
Haupteffekte und Interaktionseffekt von Concept Map-Aufgabenformat und –Bewertungsbogen	parametrisch: Zweifaktorielle ANOVA und Profilplots		Gruppenzugehörigkeit der Lehrerinnen und Lehrer

5 Ergebnisse und Hypothesenprüfung

Im Ergebnisteil werden für die jeweiligen Studien zu Beginn die deskriptiven Statistiken aufgeführt. Es folgen die Ergebnisse zur Prüfung der Analysevoraussetzungen und abschließend die Hauptergebnisse der jeweiligen Studie hinsichtlich der Hypothesen.

5.1 Studie 1

Ziel dieser Studie ist es festzustellen, welche Konstrukte Concept Maps messen. Es werden die entwickelten Concept Map-Aufgabenformate und das Concept Map-Bewertungsformat gegen einen Kompetenztest konvergent validiert.

5.1.1 Deskriptive Ergebnisse

Die N = 79 Schülerinnen und Schüler stammen aus vier 9.Klassen zweier G8-Gymnasien in Nordrhein-Westfalen. Die Datenerhebung wird für jede Klasse an zwei Tagen durchgeführt. Insgesamt bearbeiten von der Gesamtstichprobe 40 Schülerinnen und Schüler das Concept Map-Aufgabenformat A (N_A = 40) und 39 Schülerinnen und Schüler das Aufgabenformat B (N_B = 39). Da diese Studie nicht das Ziel verfolgt, Geschlechterunterschiede zwischen den Schülerinnen und Schülern, kognitive Fähigkeiten und das Alter aufzuschlüsseln, werden diese Variablen nicht erhoben. Die Anzahl der von den Schülerinnen und Schülern richtig gelösten Aufgaben im Kompetenztest beträgt im Mittel M = 8.84, SD = 3.48. Es können maximal 22 Aufgaben richtig beantwortet werden.

5.1.2 Ergebnisse zur konvergenten Validität

Voraussetzungen und Datenaufbereitung

Doppelkodierung und Normalverteilung

Um Fehler bei der Dateneingabe des Kompetenztests auszuschließen, werden 10% aller vorliegenden Testhefte doppelt eingegeben und die Interraterübereinstimmung dieser Eingabe über Cohens κ berechnet. Statt 7,9 Testhefte werden 10 Testhefte doppelt kodiert, aufgeschlüsselt in insgesamt 220 Eingaben. Es ergibt sich für diese Übereinstimmungsprüfung ein κ-Wert von .99. Nach Bortz und Döring (2006) erfordert eine zufriedenstellende Übereinstimmung κ-Werte von mindestens .60.

Durch den Kolmogorov-Smirnov-Test für eine Stichprobe wird geprüft, inwieweit die Schülerleistung im Kompetenztest normalverteilt ist. Die Prüfung zeigt, dass sich die Stichprobenwerte hinsichtlich dieses Merkmals signifikant normalverteilen (D (79) = .09, p = .08) (siehe Abbildung 5.1).

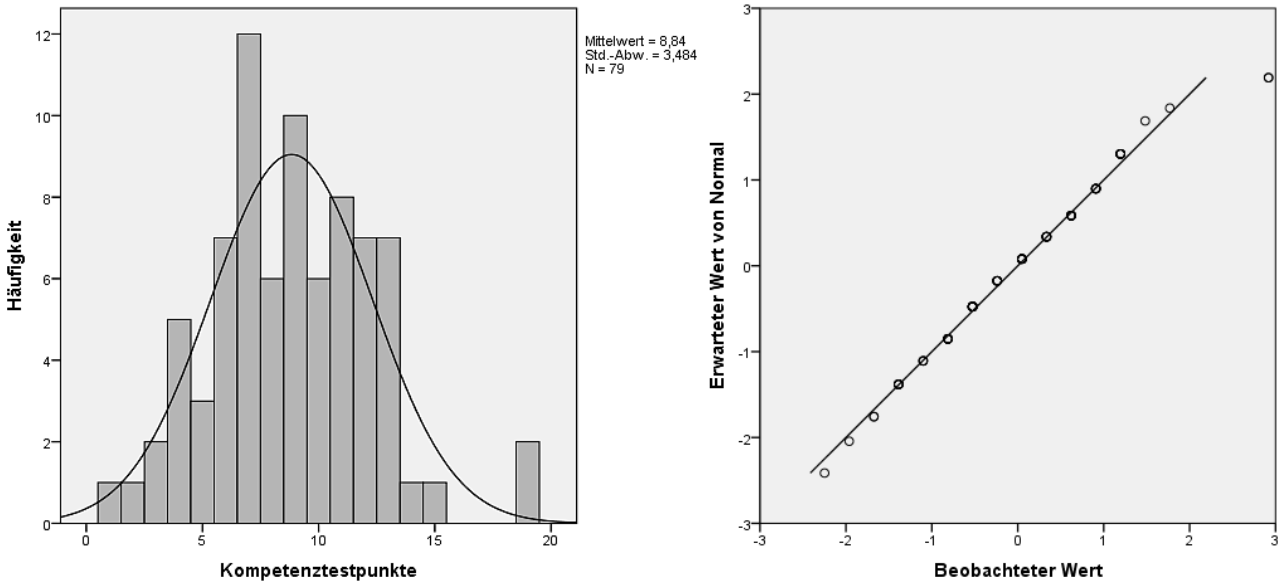


Abbildung 5.1. Links: Histogramm der Schülerstichprobe im Kompetenztest.
Rechts: Q-Q-Normalverteilungsdiagramm der z-standardisierten Kompetenztestpunkte.

Bemerkungen: Im Q-Q-Normalverteilungsdiagramm wird der beobachtete Wert im Kompetenztest gegen den Wert aufgetragen, der bei einer Normalverteilung erwartet werden kann. Liegen die Punkte im Q-Q-Normalverteilungsdiagramm auf der Geraden, kann von einer Normalverteilung ausgegangen werden. Dies ist der Fall.

Zusätzlich wird mit dem Levene-Test geprüft, ob sich die Varianzen der Gruppen mit den unterschiedlichen Aufgabenformaten hinsichtlich der Kompetenztestleistung homogen verhalten. Die Varianzen sind in den beiden Schülergruppen nicht signifikant unterschiedlich, F (1,77) = 1.34, p = .25.

Die Testvoraussetzungen Normalverteilung und Varianzhomogenität sind erfüllt, sodass die Analysen für diese Studie parametrisch durchgeführt werden können.

Im Folgenden werden die Analysewege erläutert. Wie in Abschnitt 4.1.4 beschrieben, erstellen die Schülerinnen und Schüler Concept Maps (mit Aufgabenformat A oder B). Anschließend werden alle 79 Concept Maps von sechs studentischen Mitarbeitern über den Concept Map-Beurteilungsbogen bewertet. Bei den Beurteilern handelt es sich um Physik-Lehramtsstudierende des Hauptstudiums für das Lehramt an Grund-, Haupt-,

Real- und Gesamtschulen mit dem Schwerpunkt Haupt-, Real- und Gesamtschule der entsprechenden Jahrgänge. Mit diesem Verfahren wird gewährleistet, dass für jede Concept Map sechs Beurteilungen vorliegen. Für alle sich anschließenden Analysen wird für jede Concept Map und somit für jedes Item, das im Beurteilungsbogen nunmehr sechsmal vorliegt, der Mittelwert für dieses Item hinsichtlich dieser einen speziellen Concept Map ermittelt. Dieser Wert ist die Grundlage für alle weiteren Berechnungen. Hinsichtlich der Leistung der Schülerinnen und Schüler im Kompetenztest wird die Summe der richtig beantworteten Aufgaben bestimmt. Dies lässt Aussagen zum allgemeinen Leistungsstand der Schülerinnen und Schüler im Konzept Energie zu (siehe Abschnitt 4.1.3 Kompetenztest (adaptiert)).

Analyseergebnisse

Interraterreliabilität und Interne Konsistenz

Die Interraterreliabilität liegt mit $ICC_{just,\ M6} = .52$ und einem Signifikanzniveau von $p < .01$ im eingeschränkt akzeptablen Bereich. Zusätzlich ermöglicht das sechsfach-Rating die Berechnung von Cronbachs α im Concept Map-Beurteilungsbogen. Mit diesem Maß kann die Skala des Bogens in der Messung des Konstrukts ‚Kompetenz im Bereich Energie' geprüft werden. Cronbachs α ergibt einen zufriedenstellenden Wert ($\alpha = .69$). Es kann davon ausgegangen werden, dass mit diesem Instrument das Konstrukt gemessen werden kann. Der Kompetenztest erfasst das Konstrukt ‚Kompetenz' eingeschränkt zufriedenstellend (Cronbachs $\alpha = .61$).

Zeitökonomie in Concept Map Beurteilung

Neben der Beurteilung der Concept Maps durch die sechs Beurteiler wird zusätzlich die Zeit ermittelt, die die jeweiligen Beurteiler für die Bearbeitung der 79 Concept Maps benötigen. Tabelle 5.1 schlüsselt die ermittelten Werte für die einzelnen Beurteiler und im Durchschnitt auf. Die Beurteilungszeit für 79 Concept Maps beträgt pro Person zwischen 300 und 660 Minuten bzw. zwischen 3.79 und 8.35 Minuten pro Map. Im Durchschnitt werden 432.5 Minuten für die Gesamtbeurteilung und 5.74 Minuten für eine Map-Beurteilung benötigt.

Tabelle 5.1. Benötigte Zeit für die Beurteilung von 79 Maps der sechs Beurteiler.

Beurteiler	Gesamtzeit für Beurteilung von 79 Maps (min)	Beurteilungszeit pro Map (min)
1	405	5.12
2	300	3.79
3	360	4.55
4	660	8.35
5	420	5.31
6	450	5.69
Durchschnitt	432.5	5.74

Konvergente Validität

Zur Bestimmung der konvergenten Validität des Concept Map-Beurteilungsbogen, angewandt auf ein spezifisches Concept Map-Aufgabenformat, wird das Ergebnis der Schülerinnen und Schüler für die jeweilige Konstellation mit den Ergebnissen des bereits validierten Kompetenztests mit der Produkt-Moment-Korrelation von Pearson korreliert. Unabhängig vom Aufgabenformat liegt der Zusammenhang zwischen Kompetenztest und Beurteilungsbogen im unteren positiven Bereich ($r = .29^*$, $p < .05$). Wird das Ergebnis für die beiden Aufgabenformate differenziert betrachtet, zeigt sich, dass die Korrelationen geringfügig steigen. Tabelle 5.2 fasst die Ergebnisse zusammen.

Tabelle 5.2. Korrelation nach Pearson zwischen Kompetenztest und Concept Map-Aufgabenformat allgemein, A und B.

Kompetenztest und Beurteilungsbogen	Kompetenztest und Beurteilungsbogen bei Aufgabenformat A	Kompetenztest und Beurteilungsbogen bei Aufgabenformat B
$r = .29^*$, $p < .05$	$r_A = .34^*$, $p < .05$	$r_B = .38^*$, $p < .05$

Bemerkung: Signifikante Ergebnisse mit 5%iger Irrtumswahrscheinlichkeit werden mit * markiert ($p < .05$). 1%ige Irrtumswahrscheinlichkeit wird mit ** markiert ($p < .01$).

Es kann aus den Ergebnissen gefolgert werden, dass Concept Maps mit dem verwendeten Aufgaben- und Bewertungsformat, Kompetenzen, wie sie im Kompetenztest gemessen werden, partiell abbilden. Hypothese 1.1, die von einer positiven Korrelation zwischen der Concept Map-Bewertung und dem Kompetenztest ausgeht, kann akzeptiert werden. Die vermutete höhere Korrelation zwischen der Concept Map-Bewertung in Aufgabenformat A und dem Kompetenztest kann nicht gezeigt werden (Hypothese 1.2). Hypothese 1.2 sollte auf Basis der geringen Korrelationsdifferenz von 0.04 zwischen Aufgabenformat A und B nicht vollends abgelehnt werden.

Gruppenunterschiede im Concept Mapping

Um Physiklehrerinnen und Physiklehrern in der zweiten Studie ein Aufgabenformat und das Bewertungsformat als praxistauglich und effizient für die Diagnose von Schülerinnen und Schülern anbieten zu können, muss eine Entscheidung für ein Aufgabenformat getroffen werden. Dies ist offen. Um zu explorieren, welches Aufgabenformat die inhaltlich gehaltvolleren Concept Maps für eine Diagnose generieren kann, wird für die zwei Schülergruppen, die die unterschiedlichen Aufgabenformate bearbeiten, der *t*-Test für unabhängige Stichproben berechnet. Die Mittelwerte der im Concept Map-Beurteilungsbogen mit Aufgabenformat A erreichten Punkte werden mit dem Ergebnis für Aufgabenformat B verglichen. Die Analyse wird durch den Mittelwertvergleich der jeweils erreichten Punkte im Kompetenztest ergänzt.

Der *t*-Test für unabhängige Stichproben zeigt, abhängig vom Concept Map-Aufgabenformat, keinen signifikanten Unterschied in der Kompetenztestleistung (t (77) = 1.07, p = .29, d = .24). Das für diese Studie interessantere Ergebnis zeigt sich in Bezug auf die Concept Map-Bewertungen. Concept Maps der Schülerinnen und Schüler, die das Aufgabenformat B bearbeiten, werden von den sechs Beurteilern höher bewertet als die Concept Maps bezüglich Aufgabenformat A (t (77) = 3.20, p < .01). Es kann angenommen werden, dass mit Aufgabenformat B inhaltlich umfangreichere Concept Maps erstellt werden. Dadurch können die Wissensstrukturen und Konzeptvorstellungen in diesem Aufgabenformat im Vergleich zur Schülergruppe mit Aufgabenformat A einfacher diagnostiziert werden. Der Effekt, der zwischen diesen beiden Schülergruppen besteht, liegt bei d = .72 und ist als starker Effekt nach Cohen (1988) einzuschätzen. Die Teststärke ist mit 1-β = .88 zufriedenstellend (vgl. Bühner & Ziegler, 2010). In Tabelle 5.3 werden die Ergebnisse des *t*-Tests dargestellt.

Tabelle 5.3. Gruppenvergleich im *t*-Test für unabhängige Stichproben.

		Aufgabenformat A^{+}	**Aufgabenformat B**$^{++}$
alle Rater	**Mittelwert der Punkte im Kompetenztest**	M = 9.25, SD = 3.82	M = 8.41, SD = 3.08
	***t*-Test**	t (77) = 1.07, p = .29, d = .24, 1-β = .18	
	Mittelwert der Punkte im CM-BB	M = 7.03, SD = 3.86	M = 9.59, SD = 3.21
	***t*-Test**	t (77) = -3.20, p < .01, d = .72, 1-β = .88	

Bemerkungen: Die Analyse wird zwischen den Schülergruppen, die das Aufgabenformat A und Aufgabenformat B bearbeitet haben, durchgeführt. $^{+}N_A$ = 40 Schülerinnen und Schüler, $^{++}N_B$ = 39 Schülerinnen und Schüler.

Computergestützte Auswertung von Concept Maps

Die Software AKOVIA (Ifenthaler, 2010) generiert für jede Concept Map zwei semantische Parameter, die den inhaltlichen Charakter der Concept Maps darstellen sollen und eine Bewertung zulassen. Diese werden in einem nächsten Schritt mit den Beurteilungen aus dem Concept Map-Beurteilungsbogen und den Punkten aus dem Kompetenztest für eine Triangulation korreliert. Dadurch sollen weitere Aussagen zur konvergenten Validität der untersuchten Concept Maps ermöglicht werden.

Die Ergebnisse werden an dieser Stelle nicht präsentiert, da die Zuverlässigkeit kritisch hinterfragt werden muss. Die Korrelationen weisen die Tendenz auf, dass alleinig zwischen dem semantischen Parameter Conceptual Matching und dem Beurteilungsbogen ein Zusammenhang besteht. Es scheint kein zusätzlicher Zusammenhang zwischen dem zweiten Parameter Propositional Matching, dem Beurteilungsbogen und dem Kompetenztest zu geben. Diese Tendenzen sollten jedoch nicht als empirische Evidenz hinsichtlich einer konvergenten Validität ausgelegt werden. Es ist unklar, wie exakt die Software die für die Generierung der semantischen Parameter notwendige Modalmap erstellt. Bei der Durchsicht der digitalen Maps treten beispielsweise Rechtschreibfehler und Wortdoppelungen auf, die die Software nicht berücksichtigt. Wenn diese Fehler nicht manuell beseitigt werden, fließen diese in die Erstellung der Modalmap ein, sodass die Vergleichsbasis für die Generierung der semantischen Parameter verzerrt ist. Die anschließenden Korrelationsberechnungen würden zu Aussagen führen, die empirisch nicht haltbar sind.

Für eine einwandfrei funktionierende automatisierte Concept Map-Analyse mittels PC müssen die Regeln der Analyse modifiziert werden. Eine Interpretation der Ergebnisse

hinsichtlich Beurteilerunabhängigkeit und konvergenter Validität ist auf dieser Grundlage nicht möglich. Die Analyse der Concept Maps mittels der Software AKOVIA wird in dieser Untersuchung nicht weiter berücksichtigt.

Aus Gründen der Untersuchungsdurchführung soll den Lehrkräften *ein* Aufgabenformat angeboten werden. Die Ergebnisse zur konvergenten Validität und der inhaltlichen Qualität der Concept Maps führen zu einer Entscheidung zu Gunsten des Aufgabenformats B.

Lehrkräften soll der Beurteilungsbogen als zeitökonomisches Instrument angeboten werden. Die Ergebnisse zur Dauer der Concept Map-Beurteilung zeigen, dass die Concept Map-Bewertung einer gesamten Klasse länger dauert als die Korrektur eines Physiktests. Das Ziel dieses Projektes ist es, den Lehrerinnen und Lehrern zeitökonomische Instrumente anzubieten. Eine Kürzung des Bogens von 18 auf 10 Items erscheint angemessen, um die Bewertung einer Concept Map nicht länger werden zu lassen, als die Korrektur eines Physiktests. Die inhaltliche Orientierung des Beurteilungsbogens an den vier Kompetenzentwicklungsstufen im Konzept Energie bleibt erhalten. Die Entscheidungsgrundlage für die Kürzung des Bogens waren die Ergebnisse einer in dieser Arbeit nicht weiter diskutierten Faktorenanalyse, die es ermöglicht, redundante Aussagen zu identifizieren und zu entfernen (siehe Abschnitt 4.2.3 Beschreibung der Instrumente).

Die Ergebnisse der Studie 1 und die ergänzende Erhebung in Physikleistungskursen (siehe Abschnitt 4.1.5) deuten darauf hin, dass sich eine Verlagerung der Schülerstichprobe in einen höheren Jahrgang empfiehlt. Die Einführungsphase (Klasse 10) sollte im Verhältnis zu einem 9. Jahrgang umfangreichere Concept Maps erstellen können, die eine Diagnostik in den höheren Kompetenzentwicklungsstufen (Stufe 3 und 4) ermöglicht.

Auf Basis der dargestellten Ergebnisse werden für das Ziel der Studie 2 und der Beantwortung der Forschungsfrage 2 drei entscheidende Modifikationen vorgenommen:

1. Den Physiklehrkräften wird nur das Aufgabenformat B als Diagnoseinstrument angeboten.
2. Der Concept Map-Beurteilungsbogen wird von 18 auf 10 Items gekürzt.
3. In Studie 2 wird die Einführungsphase (Klasse 10) als Zielgruppe angesprochen.

5.2 Studie 2

Aufbauend auf den in Studie 1 erprobten Instrumenten, soll die Diagnosegenauigkeit der Lehrkräfte über die Nutzung der Instrumente eingeschätzt werden.

5.2.1 Deskriptive Ergebnisse

Lehrerstichprobe

Die Studie wird mit 48 Physiklehrkräften und ihren jeweiligen Physik-Einführungs-Kursen durchgeführt. Es lässt sich ein Überhang an Physiklehrern feststellen (m = 87.50% und w = 12.5%). Randomisiert werden 13 Lehrpersonen der Gruppe 1 zugeordnet, 14 der Gruppe 2, 12 der Gruppe 3 und neun der Gruppe 4. Im Durchschnitt sind die Lehrerinnen und Lehrer 41 Jahre alt (SD = 10.65) und arbeiten im Mittel seit 11.85 Jahren (SD = 11.17) an der Schule. 11 von 47 Lehrkräfte kennen bereits die Methode des Concept Mappings, 29 von 45 lassen weder Concept Maps von den Schülerinnen und Schülern erstellen, noch erstellen sie selber Concept Maps. Die Studie konnte nicht innerhalb eines Schulhalbjahres beendet werden. Es wurde ein Teil der Stichprobe im sich anschließenden Schulhalbjahr erhoben. Die deskriptiven Statistiken zur Lehrerstichprobe und den Erhebungszeitpunkten werden in den Tabellen 5.6a und 5.6b dargestellt.

Schülerstichprobe

Insgesamt nehmen N = 977 Schülerinnen und Schüler aus Physikkursen der Einführungsphase (EF, Klasse 10) der G8-Gymnasien teil. Die Physikkurse werden von den jeweiligen Physiklehrerinnen und Physiklehrern, die an dieser Studie teilnehmen, unterrichtet, sodass die jeweiligen Physikkurse der gleichen Gruppe zugeordnet werden wie die Lehrkraft (z. B. Lehrer X gehört der Gruppe 1 an, sein Kurs ebenfalls der Gruppe 1). Der KFT wird von N = 971 Schülerinnen und Schülern bearbeitet (sechs Personen haben an diesem Messzeitpunkt gefehlt). Es nehmen mehr Schüler als Schülerinnen teil (m = 65.30% und w = 33.70%). Dies lässt sich mit der Kurswahl der Schülerinnen und Schüler erklären, die vor Eintritt in die Einführungsphase geschehen muss. In der Einführungsphase gibt es keinen Physikunterricht im Klassenverband.

Die Tabellen 5.4, 5.5a und 5.5b fassen die Ergebnisse zur Schülerstichprobe zusammen. Die Stichprobenverteilung der Lehrerinnen und Lehrer in die Gruppen wird auf der Seite 73 in den Tabellen 5.6a und 5.6b dargestellt.

Tabelle 5.4. Deskriptive Statistiken für die Schülerstichprobe.

Variable	Schülerinnen und Schüler
Gesamtanzahl KT	$N = 977$
Gesamtanzahl KFT	$N = 971$
Gesamtzahl auf Gruppen verteilt	Gr. 1 = 255 Gr. 2 = 304 Gr. 3 = 239 Gr. 4 = 179
Geschlechterverteilung	$w = 33.70\%$ $m = 65.30\%$ Fehlend = 1%
Altersdurchschnitt (in Jahren)	$M_{gesamt} = 15.82$ ($SD = .79$) $M_w = 15.78$ ($SD = .76$) $M_m = 15.83$($SD = .79$)
Mittlere Leistung im Kompetenztest (Anzahl richtiger Antworten)	$M_{gesamt} = 12.21$ ($SD = 4.51$) $M_w = 10.94$ ($SD = 4.26$) $M_m = 12.90$ ($SD = 4.50$)
Mittlere Leistung im KFT (Anzahl richtiger Antworten)	$M_{gesamt} = 18.07$ ($SD = 4.41$) $M_w = 18.02$ ($SD = 4.41$) $M_m = 18.10$ ($SD = 4.41$)

Bemerkung: Im Kompetenztest konnten maximal 22 richtige Antworten gegeben werden. Im KFT konnten 25 richtige Antworten erreicht werden.

Tabelle 5.5a. Mittlere Schülerleistung im Kompetenztest (KT) in Abhängigkeit von der Gruppe.

Gruppe	*MW* der erreichten Punkte im KT	*N*	*SD*
1	12,18	255	4,77
2	12,52	304	4,39
3	11,03	239	4,28
4	13,33	179	4,29
Insgesamt	12,21	977	4,51

Bemerkung: Im Kompetenztest konnten maximal 22 richtige Antworten gegeben werden.

Tabelle 5.5b. Mittlere Schülerleistung im KFT in Abhängigkeit von der Gruppe.

Gruppe	*MW* der erreichten Punkte im KFT	*N*	*SD*
1	18,18	250	4,41
2	18,68	304	3,91
3	17,79	238	4,56
4	17,22	179	4,87
Insgesamt	18,07	971	4,41

Bemerkung: Im KFT konnten 25 richtige Antworten erreicht werden.

Tabelle 5.6a. Deskriptive Statistiken für die Lehrerstichprobe.

Variable	Lehrerinnen und Lehrer
Gesamtanzahl	$N = 48$
Gesamtzahl auf Gruppen verteilt	Gr. 1 = 13 Gr. 2 = 14 Gr. 3 = 12 Gr. 4 = 9
Geschlechtsverteilung gesamt	$w = 6$ (12.50%) $m = 42$ (87.50%)
Gruppengeschlechtsverteilung	Gr. 1: $w = 4$, $m = 9$ Gr. 2: $w = 2$, $m = 12$ Gr. 3: $w = 0$, $m = 12$ Gr. 4: $w = 0$, $m = 9$
Altersdurchschnitt	$M_{gesamt} = 41.64$ ($SD = 10.65$) $M_w = 38.50$ ($SD = 10.69$) $M_m = 42.10$ ($SD = 10.70$)
Durchschnittliche Berufsjahre	$M_{gesamt} = 11.85$ ($SD = 11.17$) $M_w = 8.16$ ($SD = 10.32$) $M_m = 12.41$ ($SD = 11.31$)
Kenntnis von Concept Maps	11 von 47 Lehrerinnen und Lehrer kennen Concept Maps
Nutzung von Concept Maps	29 von 45 Lehrerinnen und Lehrer nutzen Concept Maps nie

Tabelle 5.6b. Deskriptive Statistiken für die Lehrerstichprobe detailliert betrachtet für die Erhebungszeitpunkte.

Ebene	Gruppe	Zeitpunkt	Anzahl	Anzahl gesamt
Lehrerinnen und Lehrer	1	2011/12	11	13
		2012/13	2	
	2	2011/12	10	14
		2012/13	4	
	3	2011/12	2	12
		2012/13	10	
	4	2011/12	9	9
		2012/13	0	

Bemerkung: Die Lehrerstichprobe musste in zwei Schulhalbjahren besucht werden: 2011/12 und 2012/13.

Zeitökonomie

Die Lehrerinnen und Lehrer der Gruppen 1 und 2 sollen zusätzlich notieren, wie viel Zeit sie für die Bewertung der Concept Maps benötigen. Die Lehrkräfte der Gruppe 3 sollen angeben, wie lange sie bei der Beurteilung mittels Beurteilungsbogen brauchen. Gruppe 4 soll keine Zeitangaben machen. Insgesamt melden 18 von 39 Lehrpersonen Zahlenwerte zurück, davon 16 der Gruppen 1 und 2 und zwei der Gruppe 3. Die Werte sind in Tabelle 5.7 dargestellt.

Tabelle 5.7. Benötigte Zeit für die Beurteilung von Concept Maps und des Beurteilungsbogens in den Gruppen 1, 2 und 3.

LuL	Beurteilungszeit pro Map/Bogen (Minimum…Maximum in min)	Mittelwert
Gruppe 1 (N = 7 von 13 LuL)	0.43 … 5.91	3.81 (SD = 1.65)
Gruppe 2 (N = 9 von 14 LuL)	0.74 … 5.77	3.38 (SD = 1.63)
Gruppe 3 (N = 2 von 12 LuL)	2.73 … 3.88	3.30 (SD = .82)
N = 18 von 39 LuL		3.55 (SD = 1.53)

5.2.2 Ergebnisse zur Diagnosegenauigkeit von Physiklehrkräften

Voraussetzungen und Datenaufbereitung

Doppelkodierung, Normalverteilung und Varianzhomogenität

Für die Prüfung einer korrekten Dateneingabe werden die Testhefte von 101 Schülerinnen und Schüler doppelt eingegeben und das Übereinstimmungsmaß κ bestimmt. Dies entspricht ca. 10% der Gesamtstichprobe mit 2.727 Eingaben für den Kompetenztest und 2.525 Eingaben für den KFT. Die Übereinstimmung für beide Tests ist ausgezeichnet, sie liegt bei κ = .99 für den Kompetenztest und bei κ = 1.00 für den KFT.

Die Leistung der Gesamtschülerstichprobe im Kompetenztest und KFT wird auf Normalverteilung überprüft. Der Kolmogorov-Smirnov-Test zeigt, dass die Schülerleistung in beiden Tests keiner Normalverteilung folgt ($D_{Kompetenztest}$ (977) = .09, p < .00 und D_{KFT} (971) = .12, p < .00, vgl. Abbildung 5.2a und 5.2b).

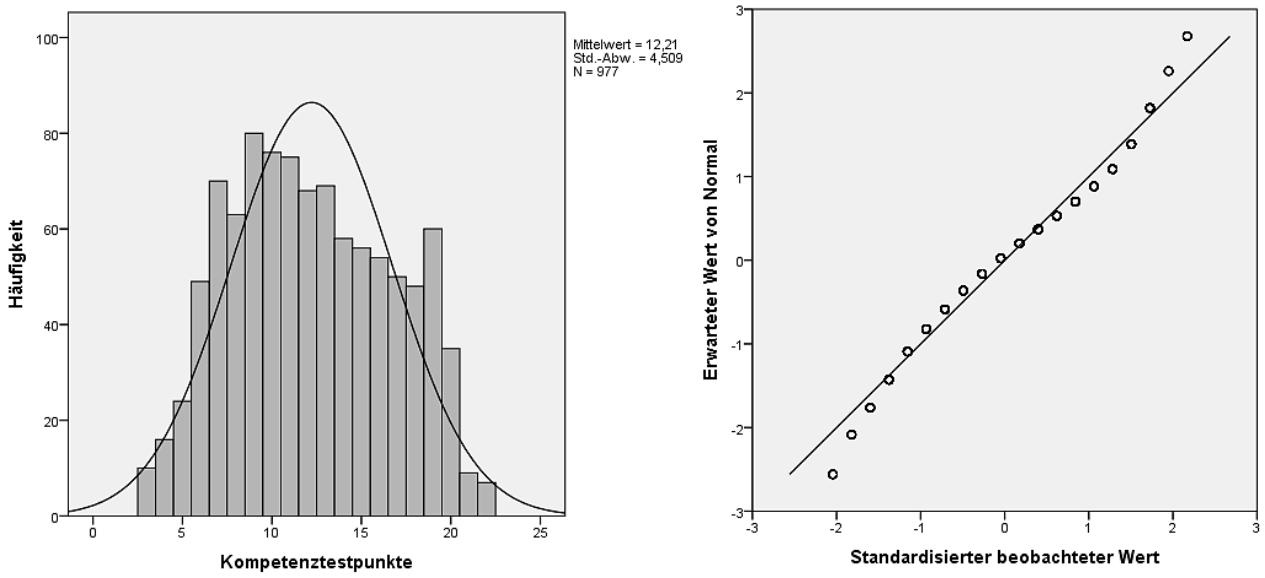


Abbildung 5.2a. Links: Histogramm der Schülerstichprobe im Kompetenztest.
Rechts: Q-Q-Normalverteilungsdiagramm der z-standardisierten Kompetenztestpunkte.

Bemerkung: Im Q-Q-Normalverteilungsdiagramm befinden sich nicht alle Punkte auf der Geraden, eine leichte Schiefe der Normalverteilung ist vorhanden.

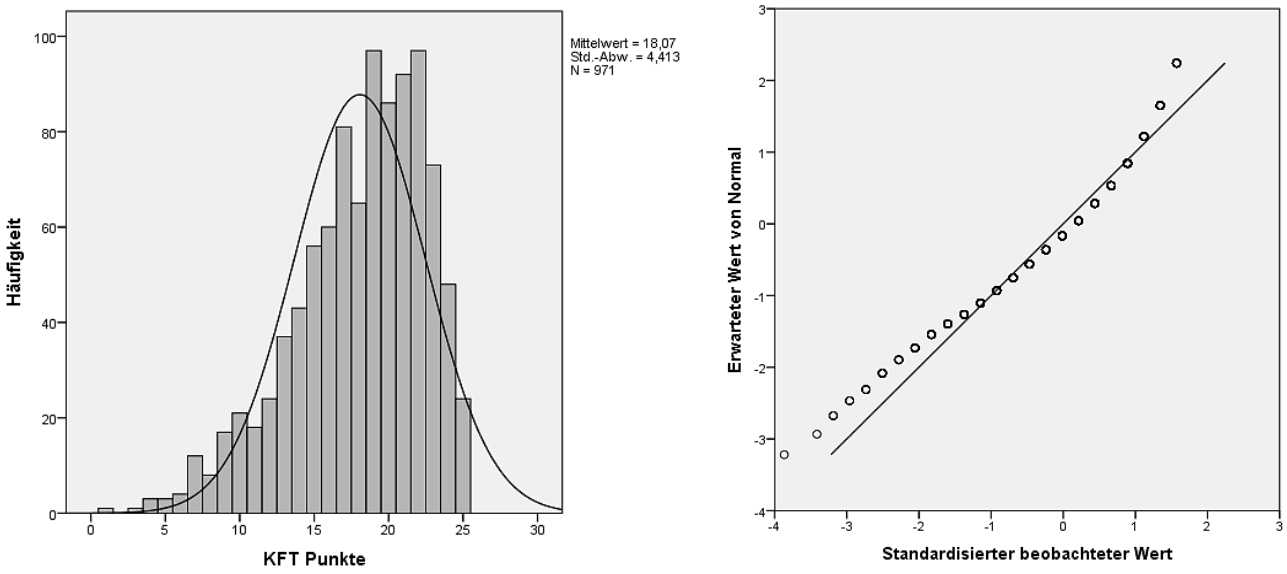


Abbildung 5.2b. Links: Histogramm der Schülerstichprobe im KFT.
Rechts: Q-Q-Normalverteilungsdiagramm der z-standardisierten KFT-Punkte.

Bemerkung: Im Q-Q-Normalverteilungsdiagramm befinden sich nicht alle Punkte auf der Geraden, eine leichte Schiefe der Normalverteilung ist vorhanden.

Die Testvoraussetzung Normalverteilung ist für die Schülerstichprobe nicht gegeben. Alle künftigen Analysen dieser Studie, die in Bezug mit der Schülerleistung im Kompetenztest und im KFT stehen, werden deshalb nicht-parametrisch durchgeführt.

Die Diagnosegenauigkeit der Physiklehrkräfte wird ebenfalls auf Normalverteilung untersucht. Hierzu werden die Rangkorrelationswerte aller Lehrkräfte, die der Diagnosegenauigkeit entsprechen, im Kolmogorov-Smirnov-Test geprüft. Die Leistung der Lehrerinnen und Lehrer weicht signifikant von normal ab (D (48) = .14, $p < .05$), wie in Abbildung 5.3 zu erkennen ist. Es wird von keiner normalverteilten Diagnosegenauigkeit ausgegangen.

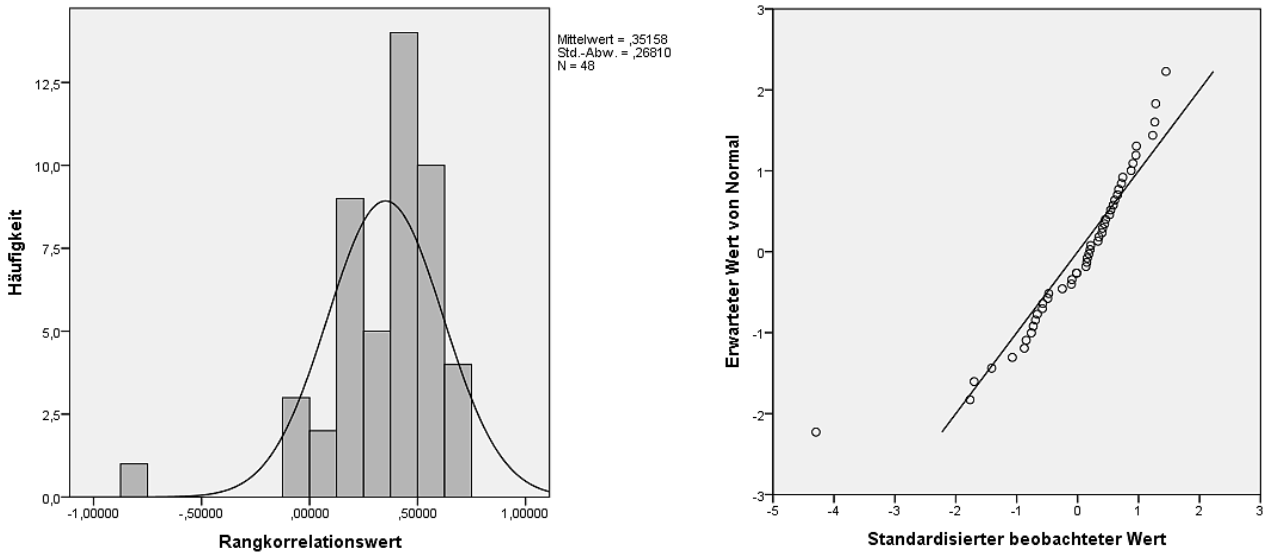


Abbildung 5.3. Links: Histogramm der Rangkorrelationswerte (Diagnosegenauigkeit) der Physiklehrkräfte. Rechts: Q-Q-Normalverteilungsdiagramm der z-standardisierten Rangkorrelationswerte.

Bemerkung: In beiden Diagrammen ist eine Schiefe zu erkennen.

Vor Beginn der Datenanalyse muss zusätzlich geprüft werden, ob zwischen den Versuchsgruppen bezogen auf die Diagnosegenauigkeit Varianzhomogenität besteht. Um dies zu prüfen, wird der Levene-Test berechnet. Die Varianzen sind in allen vier Gruppen nicht signifikant unterschiedlich, F (3, 44) = 2.72, p = .056.

Nach den genannten Testvoraussetzungen für die Diagnosegenauigkeit der Lehrerstichprobe werden alle künftigen statistischen Tests nicht-parametrisch durchgeführt.

Analyseergebnisse

Interne Konsistenz

Die Skala des Kompetenztests wird in dieser Studie erneut auf seine interne Konsistenz überprüft. Cronbachs $\alpha =$ .80 kann weiterhin als gut eingestuft werden. Zusätzlich wird der KFT hinsichtlich der Erfassung des Konstrukts kognitive Fähigkeiten untersucht. Cronbachs α liegt für diese Skala bei .79.

Der (Concept Map-) Beurteilungsbogen wird in den Gruppen 1 und 3 eingesetzt. Für jede einzelne Lehrperson, die den Bogen nutzt, kann errechnet werden, wie konsistent der Bogen das Konstrukt ‚Kompetenz im Bereich Energie' misst. Ausgehend von diesen individuellen Werten wird für die jeweiligen Gruppen Cronbachs α gemittelt errechnet. Für Gruppe 1 (N_{LuL} = 13) ist das gemittelte α akzeptabel (Cronbachs $\alpha_{\text{Gruppe 1, mittel}}$ = .66). Die interne Konsistenz des Bogens kann für Gruppe 3 (N_{LuL} = 12) als zufriedenstellend eingeschätzt werden (Cronbachs $\alpha_{\text{Gruppe 3, mittel}}$ = .92).

Alle drei Instrumente erfassen die jeweiligen Konstrukte.

Einfluss der Gruppenzugehörigkeit auf die Diagnosegenauigkeit

Alle Lehrkräfte müssen eine Rangfolge ihrer Schülerinnen und Schüler erstellen. Diese Rangfolge wird mit der Rangfolge der Schülerinnen und Schüler aus dem Kompetenztest als Spearmans Rangkorrelation verglichen. Jede Lehrerin und jeder Lehrer erhält auf diese Weise einen Rangkorrelationswert, der als ein Maß für die Diagnosegenauigkeit der betreffenden Lehrperson angesehen werden kann. Die Rangkorrelationen bewegen sich für die einzelnen Versuchsgruppen in verschiedenen Bereichen. Die Tabellen 5.8a, 5.8b und 5.8c zeigen die Intervalle und die Mediane für die Gruppen und differenziert für die Geschlechter.

Tabelle 5.8a. Spearmans Rangkorrelationen für die einzelnen Gruppen.

	Gruppe 1 (*N* = 13)	**Gruppe 2 (*N* = 14)**	**Gruppe 3 (*N* = 12)**	**Gruppe 4 (*N* = 9)**
Spearmans Rangkorrelation ρ	ρ = -.12 … .69	ρ = -.80 … .74	ρ = .13 … .70	ρ = .40 … .59
Median (*Md*) $\text{Median}_{\text{gesamt}}$ = .403	.31	.22	.43	.52

Bemerkungen: Der Median wird aufgeführt, da die Diagnosegenauigkeit keiner Normalverteilung folgt. Jeder Lehrkraft wird auf Basis seines Rangkorrelationswerts eine Rangposition zugeordnet. Diese dient als Basis für alle weiteren Berechnungen.

Tabelle 5.8b. Spearmans Rangkorrelationen für die einzelnen Gruppen hinsichtlich der Lehrerinnen.

	Gruppe 1 (*N* = 4)	**Gruppe 2 (*N* = 2)**	**Gruppe 3 (*N* = 0)**	**Gruppe 4 (*N* = 0)**
Spearmans Rangkorrelation ρ	ρ = -.12 … .44	ρ = .15 … .77	ρ = -	ρ = .-
Median (*Md*) $Median_{gesamt}$ = .211	.21	.31	-	-

Tabelle 5.8c. Spearmans Rangkorrelationen für die einzelnen Gruppen hinsichtlich der Lehrer.

	Gruppe 1 (*N* = 9)	**Gruppe 2 (*N* = 12)**	**Gruppe 3 (*N* = 12)**	**Gruppe 4 (*N* = 9)**
Spearmans Rangkorrelation ρ	ρ = -.03 … .69	ρ = -.80 … .74	ρ = .13 … .70	ρ = .40 … .59
Median (*Md*) $Median_{gesamt}$ = .407	.35	.25	.43	.52

Inwiefern sich die Diagnosegenauigkeit der Gruppen voneinander unterscheidet und inwiefern die Mediane als statistisch bedeutsam eingestuft werden können, lässt sich über die Rangvarianzanalyse nach Kruskal und Wallis (*H*-Test) (vgl. Field, 2009) feststellen.

Es kann ein genereller Einfluss der Gruppe, in der sich eine Lehrkraft befindet, auf die Diagnosegenauigkeit verzeichnet werden (H (3) = 10.78, $p < .05$). Die Stärke dieses Effektes kann mit ω = .47 als moderater Effekt aufgefasst werden. Für detaillierte Gruppenpaarvergleiche mittels Post-Hoc-*U*-Tests wird das kritische Signifikanzlevel für diese Vergleiche von p = .05 auf .0083 nach der Bonferroni-Korrektur (vgl. Field, 2009) adjustiert. Die Gruppenvergleiche zeigen, dass zwischen den Gruppen 1 und 4, Gruppe 1 und 3 und 3 und 4 kein signifikanter Unterschied in der Diagnosegenauigkeit besteht. Lediglich die Gruppen 2 und 4 lassen sich in ihrer Diagnosegenauigkeit voneinander trennen.

Die Ergebnisse lassen vermuten, dass Lehrkräfte, die ihre Schülerinnen und Schüler anonym anhand einer Concept Map mit Hilfe eines Beurteilungsbogens bewerten (Gruppe 1), die Rangordnung ihrer Schülerinnen und Schüler ebenso gut bilden können, wie Lehrkräfte, die ihre Schülerinnen und Schüler personalisiert auf Basis ihrer Unterrichtsbeobachtungen einschätzen (Gruppe 4). Der Einfluss der Gruppenzugehörigkeit auf die Höhe der Diagnosegenauigkeit der Physiklehrkräfte wird in Abbildung 5.4 (*H*-Test) und Tabelle 5.9 (Post Hoc-*U*-Tests) gezeigt.

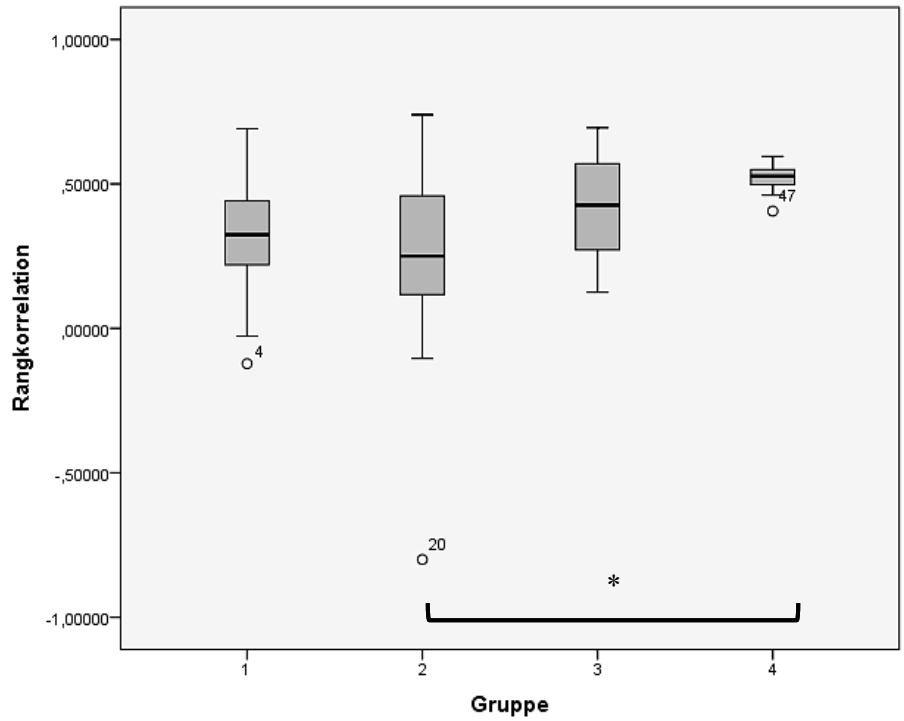


Abbildung 5.4. Boxplot der Rangkorrelationswerte (Diagnosegenauigkeit) bezogen auf die Gruppen (*H*-Test).

Bemerkung: Es werden nur die signifikanten Ergebnisse markiert. * $p < .0083$.

Tabelle 5.9. Mehrgruppenvergleiche im Post Hoc-*U*-Test bezogen auf die Diagnosegenauigkeit.

(I) Gruppe	(J) Gruppe	*U*	*z*	Signifikanz	ω	Cohens *d*	1-β
1	2	77.00	-.68	.52	.13	.30	.03
	3	57.50	-1.12	.28	.22	.57	.09
	4	21.00	-2.50	.01	.53	1.25	.50
2	3	50.00	-1.75	.09	.34	.73	.18
	4	14.00	-3.09	.001*	.64	1.16	.44
3	4	38.00	-1.14	.28	.25	.64	.09

Bemerkungen: adjustiertes Signifikanzlevel nach Bonferroni-Korrektur bei * $p < .0083$ (vgl. Field, 2009). Cohens *d* wird mit den parametrisch ermittelten Mittelwerten der Diagnosegenauigkeit der Gruppen ermittelt, um näherungsweise die Teststärke 1-β des *U*-Tests berechnen zu können. Die durch dieses Verfahren ermittelte Teststärke sollte allerdings nur als Richtwert einer Mindestteststärke aufgefasst werden. Die Nutzung des parametrischen *d* führt zu einer Unterschätzung der wahren Teststärke in nicht-parametrischen Tests (vgl. Rasch, Friese, Hofmann & Naumann, 2010).

Tabelle 5.9 zeigt zusätzlich die Effektstärken und Teststärken der *U*-Tests für die Gruppenvergleiche. Trotz hoher Effektstärke des nicht-signifikanten Paarvergleichs von Gruppe 1 und 4 mit $\omega = .53$, schränkt die Teststärke 1-$\beta = .50$ das nicht-signifikante

Ergebnis ein. Die berechnete Teststärke ist als Richtwert einer Mindestteststärke aufzufassen. Es muss angenommen werden, dass der *U*-Test mit einer Wahrscheinlichkeit von mindestens 50% einen Unterschied zwischen den Gruppen 1 und 4 aufdecken kann. Der in diesem Vergleich nicht gefundene Unterschied sollte vor diesem Hintergrund kritisch betrachtet werden.

In Hypothese H 2.1 wird angenommen, dass Physiklehrkräfte in der Lage sind, eine Rangfolge ihrer Schülerinnen und Schüler bilden zu können, die der Rangfolge eines Kompetenztests ähnlich ist. Die Ergebnisse zeigen, dass Lehrkräfte Rangordnungen ihrer Schülerinnen und Schüler bilden können und dies angemessen tun (vgl. Rangkorrelationswerte in Abbildung 5.4). Die Spannweite zwischen den einzelnen Lehrkräften ist groß, dennoch wird ein Trend deutlich, der auf eine grundsätzlich vorhandene Fähigkeit der Lehrkräfte zur Rangordnungsbildung schließen lässt. Die Hypothese H 2.1 kann akzeptiert werden. Die Höhe der Rangordnungsübereinstimmung in Form der mittleren Rangkorrelation ist in Gruppe 4 am höchsten. Dies widerspricht der Hypothese 2.2, in der bei Gruppe 1 von der höchsten Rangkorrelation ausgegangen wird. Die Lehrkräfte der Gruppe 1, in der Concept Maps und der Beurteilungsbogen eingesetzt werden, weisen im Median eine niedrigere Rangkorrelation im Vergleich zu den Gruppen 4 und 3 auf. Zusätzlich zeigt die Prüfung der statistischen Relevanz dieses Unterschieds, dass die Gruppen 1 und 4 sich jedoch nicht unterscheiden (vgl. Tabelle 5.9). Die Teststärkeberechnungen zeigen zudem, dass das Ergebnis (die Gruppen 1 und 4 können nicht getrennt werden) kritisch betrachtet werden muss. Es kann auf Basis dieser Analysen nicht davon ausgegangen werden, dass Gruppe 1 eine höhere Diagnosegenauigkeit aufweist als die Gruppen 2, 3 und 4. Gruppe 1 lässt sich möglicherweise von Gruppe 4 trennen, wenn die Teststärke berücksichtigt wird. Diese Befunde führen dazu, dass Hypothese 2.2 abgelehnt wird.

Zusammenhang der Kontrollvariablen mit der Schülerleistung im Kompetenztest und im Concept Mapping

Welcher Zusammenhang zwischen der Kompetenztestleistung der Schülerinnen und Schüler mit den kognitiven Fähigkeiten, der letzten Physiknote, Mathematiknote und Deutschnote besteht, wird durch eine Korrelationsberechnung nach Spearman statistisch aufgezeigt. Es wird davon ausgegangen, dass ein signifikanter Zusammenhang zwischen den Kontrollvariablen und der Leistung im Kompetenztest besteht. Tabelle 5.10 zeigt die Ergebnisse der Korrelationsberechnung.

Tabelle 5.10. Korrelationsberechnungen nach Spearman für die Schülerstichprobe.

		Kompetenz-test	KFT	Physik-note	Mathe-note	Deutsch-note
Kompetenztest	ρ	1				
	Signifikanz					
	N	977				
KFT	ρ	.322**	1			
	Signifikanz	.000				
	N	971	971			
Physiknote	ρ	.373**	.274**	1		
	Signifikanz	.000	.000			
	N	964	958	964		
Mathenote	ρ	.333**	.258**	.595**	1	
	Signifikanz	.000	.000	.000		
	N	967	961	961	967	
Deutschnote	ρ	.097**	.106**	.367**	.397**	1
	Signifikanz	.003	.000	.000	.000	
	N	966	960	958	963	966

Bemerkung: * $p < .05$, ** $p < .01$.

Alle relevanten Kontrollvariablen weisen unterschiedlich hohe Zusammenhänge zueinander auf. Die kognitiven Fähigkeiten korrelieren mit der Kompetenztestleistung positiv. Die Schulnoten korrelieren ebenfalls jeweils positiv mit der Kompetenztestleistung und dem KFT.

Inwiefern die KFT-Leistung mit der Leistung in den Concept Maps der Gruppe 1 zusammenhängt, wird durch Spearmans Rangkorrelationskoeffizienten ρ exploriert. Die Punkte, die die Schülerinnen und Schüler für ihre Concept Maps im Beurteilungsbogen erhalten, werden mit der Leistung im KFT in Bezug gesetzt. Der KFT korreliert niedrig, aber signifikant mit der Concept Map-Bepunktung ($\rho = .197^{**}$, $p < .01$).

Zusammenhang der Kontrollvariablen mit der Diagnosegenauigkeit

Lehrermerkmale wie Alter und Anzahl der Berufsjahre der Lehrkräfte können neben der Gruppenzugehörigkeit ebenfalls in Zusammenhang mit der Höhe der Diagnosegenauigkeit stehen. Um dies zu explorieren, werden Korrelationen nach Spearman berechnet. Die Korrelationen werden für die Kontrollvariablen Anzahl der Berufsjahre, Alter, Geschlecht der Lehrkräfte, die Kenntnis von Concept Maps und die Nutzung von Concept Maps bezogen auf die Diagnosegenauigkeit erstellt. Tabelle 5.11 stellt die Zusammenhänge dar.

Tabelle 5.11. Korrelationen nach Spearman für die Lehrerstichprobe mit Lehrermerkmalen.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Diagnose-genauigkeit	ρ	1						
	Signifikanz							
	N	48						
(2) Gruppe	ρ	.394**	1					
	Signifikanz	.006						
	N	48	48					
(3) Anzahl der Berufsjahre	ρ	-.207	-.060	1				
	Signifikanz	.159	.687					
	N	48	48	48				
(4) Alter	ρ	-.490	-.044	.872**	1			
	Signifikanz	.745	.771	.000				
	N	47	47	47	47			
(5) Geschlecht	ρ	.232	.367*	.105	.118	1		
	Signifikanz	.113	.010	.478	.431			
	N	48	48	48	47	48		
(6) Kenntnis von CM	ρ	-.120	-.180	.204	.254	-.061	1	
	Signifikanz	.420	.225	.168	.089	.684		
	N	47	47	47	46	47	45	
(7) Nutzung von CM	ρ	.137	-.003	-.140	-.027	-.006	-.294	1
	Signifikanz	.370	.983	.360	.860	.968	.052	
	N	45	45	45	45	45	44	45

Bemerkungen: Die Variable ‚Kenntnis von CM' fragt die Lehrerinnen und Lehrer, ob sie Concept Mapping bereits kennen. Die Variable ‚Nutzung von CM' fragt die Lehrerinnen und Lehrer, inwiefern sie Concept Maps nutzen. Die genauen Fragen können im Anhang eingesehen werden. * $p < .05$, ** $p < .01$.

Zwischen der Gruppenzugehörigkeit einer Lehrkraft und der Diagnosegenauigkeit besteht ein mittelhoher signifikanter Zusammenhang. Dies deckt sich mit den bereits zuvor dargestellten Ergebnissen zum Einfluss der Gruppenzugehörigkeit auf die Diagnosegenauigkeit. Diese Korrelation ist nicht unerwartet. Eine differenzierte Betrachtung hinsichtlich der Geschlechter zeigt, dass die Lehrerinnen im Median eine niedrigere Diagnosegenauigkeit aufweisen als die Lehrer (vgl. S. 79). Jedoch ist dieser Unterschied statistisch nicht signifikant ($U = 75$, $p = .112$, $z = -1.60$, $\omega = .23$).

Die Kontrollvariablen auf Lehrerseite stehen nicht in Zusammenhang mit der Diagnosegenauigkeit.

Wenn die Lehrermerkmale keinen Zusammenhang mit der Höhe der Diagnosegenauigkeit haben, stellt sich die Frage, ob die Klassenstruktur, d. h. die Leistung der Schülerinnen und Schüler im Kompetenztest und im KFT, mit der Höhe der

Diagnosegenauigkeit der Lehrkräfte in Zusammenhang steht. Hierzu wird die mittlere Leistung pro Klasse im Kompetenztest und KFT ermittelt und mit der Diagnosegenauigkeit der entsprechenden Lehrkraft nach Spearman korreliert (s. Tabelle 5.12).

Tabelle 5.12. Korrelationsberechnungen nach Spearman für die gesamte Lehrerstichprobe mit Klassenstrukturmerkmalen.

		(1)	(2)	(3)
(1) Diagnosegenauigkeit	ρ	1		
	Signifikanz			
	N	48		
(2)Durchschnittsleistung im Kompetenztest	ρ	.127	1	
	Signifikanz	.388		
	N	48	48	
(3) Durchschnittsleistung im KFT	ρ	-.141	.506**	1
	Signifikanz	.340	.000	
	N	48	48	48

Bemerkung: Für jede Klasse wird die durchschnittliche Leistung im Kompetenztest und KFT ermittelt. * $p < .05$, ** $p < .01$.

Tabelle 5.12 zeigt keinen signifikanten Zusammenhang zwischen der Diagnosegenauigkeit der Lehrkräfte und der Durchschnittsleistung der Schülerinnen und Schüler im Kompetenztest und KFT. Es wird vermutet, dass die Diagnosegenauigkeit der Lehrkräfte nicht von der Schülerleistung beeinflusst wird.

6 Diskussion

Dieses abschließende Kapitel diskutiert die zuvor dargestellten Ergebnisse und weist zusätzlich auf Probleme und Grenzen der beiden Studien hin.

Studie 1

Interraterreliabilität

Die von Schülerinnen und Schülern erstellten Concept Maps (N = 79) werden von sechs verschiedenen Beurteilern mittels Concept Map-Beurteilungsbogen beurteilt. Der Wert ($ICC_{just,\ M6}$ = .52) ist vor dem Hintergrund, dass es sich um eine Feldstudie handelt, nachvollziehbar und akzeptabel.

Die Interpretation des Ergebnisses zur Beurteilerübereinstimmung muss zusätzlich berücksichtigen, dass es sich bei den Beurteilern um studentische Mitarbeiter handelt, die nicht die Erfahrungen einer Physiklehrkraft haben. Alle Beurteiler haben einen physikbezogenen Lehramtshintergrund und wurden in das Bewerten von Concept Maps eingewiesen. Diese Unterweisung fand jedoch nicht im Sinne eines intensiven Kodierertrainings mit einem detaillierten Kodiermanual statt, wie es beispielsweise bei einer Videokodierung vorgesehen ist. Die studentischen Mitarbeiter sollten das Bewertungsverfahren simulieren, wie es später von Lehrkräften durchgeführt werden sollte. Lehrerinnen und Lehrer haben im Schulalltag keine Zeit für ein intensives Bewertungstraining und das Lesen eines detaillierten Kodiermanuals. Eine intensive Schulung der studentischen Mitarbeiterinnen und Mitarbeiter und der Lehrerinnen und Lehrer sollte deshalb zur besten Einschätzung des Verfahrens ähnlich erfolgen. Der Concept Map-Beurteilungsbogen musste selbsterklärend konstruiert werden. Zusätzlich wurde eine kurze Instruktionsanleitung zur Nutzung des Bogens entwickelt.

Wie bei den studentischen Mitarbeitern ist die mittlere Interraterübereinstimmung der Physiklehrerinnen und Physiklehrer (Cohens κ = .47 bei drei Lehrerpaarvergleichen), die ebenfalls die Concept Maps bewerten, akzeptabel (siehe Abschnitt 4.1.5 Ergänzende Schritte nach Studie 1). Dies verdeutlicht, dass das gesamte Concept Mapping-Bewertungsverfahren hochinferent ist und bereits mittelhohe Beurteilerübereinstimmungen als zufriedenstellend eingeschätzt werden können.

Hypothese 1.1 und Hypothese 1.2

Die Hypothese der Studie 1 lautet: H 1.1 Es besteht eine positive Korrelation im unteren Bereich zwischen Bewertung der Concept Maps über Beurteilungsbogen und Kompetenztest.

Durch die Ergebnisse aus Kapitel 5 kann die Hypothese akzeptiert werden. Die Gesamtkorrelationsberechnung, die keinen Unterschied zwischen den Concept Map-Aufgabenformaten vornimmt und die Differenzierung für die einzelnen Aufgabenformate, zeigen Korrelationen im unteren positiven Bereich (r = .29*, p < .05, r_A = .34*, p < .05, r_B = .38*, p < .05). Dieses Ergebnis steht im Einklang mit bereits bestehenden Forschungsergebnissen, die über konvergente bzw. divergente Validitäten berichten (vgl. u. a. Übersicht in Ruiz-Primo & Shavelson, 1996). Die Spannweite der Validitäten, über die bei Ruiz-Primo und Shavelson (1996) berichtetet wird, ist groß. Das jeweilige Concept Map-Aufgabenformat und die Instrumente zur Validierung bestimmen die Höhe der Validität der jeweiligen Studie. Geschlossene Concept Map-Aufgabenformate korrelieren mit geschlossenen Testaufgaben, wie Multiple-Choice-Aufgaben, höher als offene Concept Map-Aufgabenformate mit diesen Testaufgaben. Deshalb wird in Hypothese 1.2 angenommen, dass eine höhere Korrelation zwischen Concept Map-Aufgabenformat A und dem Kompetenztest besteht als zwischen Concept Map-Aufgabenformat B und dem Kompetenztest. Die Ergebnisse zeigen, dass die Hypothese nicht akzeptiert werden kann. Die Korrelation zwischen Aufgabenformat A und dem Kompetenztest ist gering niedriger als zwischen Aufgabenformat B und dem Test, sodass sich die Korrelationen der Aufgabenformate A und B zum Kompetenztest nicht unterscheiden.

Aus den Ergebnissen ergeben sich folgende weitere Fragen:

1. Warum fallen alle Korrelationen generell nicht höher aus und

2. Warum korreliert Aufgabenformat B entgegen der Hypothese ähnlich hoch mit dem Kompetenztest wie Aufgabenformat A mit dem Test?

Zu 1. Aus der Forschung ist bereits bekannt, dass Concept Map-Aufgabenformate und Bewertungsformate Komponenten von Wissen abbilden, die mit anderen Verfahren nicht erschlossen werden können (Fischler & Peuckert, 2000). Das heißt, dass ein Instrument wie der Kompetenztest nicht das gleiche Konstrukt messen kann wie das für dieses Projekt entwickelte Concept Map-Aufgabenformat und Bewertungsformat. Viering (2012) kann zeigen, dass sein Kompetenztest die Kompetenz im Bereich Energie misst. Für das Concept Map-Verfahren muss auf Basis der Ergebnisse festgehalten werden, dass das Konstrukt ‚Kompetenz' durch das Concept Mapping nicht vollständig abgebildet

werden kann. Das Konstrukt muss umfangreicher sein, als die beiden Verfahren Concept Mapping und Kompetenztest es erfassen können. Die in diesem Projekt gefundenen Zusammenhänge erscheinen zwar niedrig, lassen sich jedoch nach bisherigen Forschungsergebnissen erwarten (vgl. u. a. Novak, Gowin & Johansen, 1983).

Zu 2. Beide Aufgabenformate unterscheiden sich offensichtlich in ihrer Anlage (vgl. die Arbeitsblätter). Aufgabenformat A ist durch die Fokussierung auf Fachbegriffe fachsprachlich orientiert, Aufgabenformat B zusätzlich anwendungs- und alltagsorientiert durch die Verbindung der Alltagsconcept-Map aus Phase 1 und den Fachbegriffen aus Phase 2. Es wurden unterschiedlich hohe Korrelationen erwartet, die nicht durch die Ergebnisse unterstützt werden können. Die ähnlich hohen Korrelationen müssen nicht zwangsläufig widersprüchlich sein. Während Aufgabenformat A hauptsächlich Wissen diagnostiziert, das als strukturell charakterisiert werden kann, ermöglicht Aufgabenformat B zusätzlich den Zugang zu kontextuell angewendetem Wissen. Die Ergebnisse deuten darauf hin, dass beide Aufgabenformate für den Kompetenztest in gleicher Weise wichtig zu sein scheinen. Aufgabenformat B korreliert mit dem Kompetenztest geringfügig höher (vgl. Tabelle 5.2), da es nicht nur die Komponenten des Aufgabenformats A misst, sondern darüber hinaus den für das Aufgabenformat B typischen eigenen Anteil.

Aufgabenformat B

Schülerinnen und Schüler, die das Concept Map-Aufgabenformat B bearbeiten, erhalten signifikant mehr Punkte in ihren Concept Maps als die Schülerinnen und Schüler mit Aufgabenformat A. Aufgabenformat B hat für die Lerner durch die Zweiphasigkeit offensichtlich mehr Möglichkeiten in der Concept Map-Ausgestaltung als Aufgabenformat A. Dies lässt darauf schließen, dass das Aufgabenformat für die Concept Map-Beurteiler umfangreicheres Material für eine Schülerdiagnose generiert. Die Schülerinnen und Schüler können in Aufgabenformat B ihre Alltagsvorstellungen präsentieren und in der zweiten Phase diese um Vorstellungen zu den Fachbegriffen erweitern. Die Concept Maps dieser Schülerinnen und Schüler ermöglichen dem Beurteiler einen leichteren Zugang zur Beurteilung und empfehlen sich für den Einsatz in der Schule.

Die in der Vorstudie entwickelten Aufgabenformate (vgl. Abschnitt 4.1.3 Concept Map-Aufgabenformat) eignen sich nicht für einen Einsatz in der Schule. Beispielsweise ist eine zeitökonomische Durchführung nicht gegeben und eine umfangreiche Schülerdiagnose durch die von den Schülerinnen und Schülern generierten Maps nicht möglich. Die Concept Maps sind in diesen Fällen für eine Diagnose nicht aussagekräftig.

Zeitökonomie

Ein Kriterium für einen praxistauglichen Einsatz von Concept Maps zur Schülerdiagnose ist die Zeitökonomie im Einsatz des Concept Map-Aufgabenformats und der anschließenden Bewertung.

Die Studie zeigt, dass inhaltlich umfangreiche Concept Maps nach einem vorangegangenen Training in 30 Minuten erstellt werden können. Lehrerinnen und Lehrer können das Aufgabenformat in ihr Stundenraster aufnehmen.

Die durchschnittliche Beurteilungszeit im Beurteilungsbogen liegt gemittelt über alle sechs Beurteiler bei 5.74 min. pro Concept Map. Mit dieser Zeit benötigt eine Lehrkraft für 20 Concept Maps (= 20 Schülerinnen und Schüler) ca. 115 Minuten. In einer Einführungsphase (Klasse 10) kann aus Erfahrungswerten mit einer Kursgröße von ca. 20 Schülerinnen und Schülern gerechnet werden. Die Korrekturzeit eines Physiktests kann bei gleicher Kursgröße als ähnlich hoch eingeschätzt werden. Die Ergebnisse zeigen ebenfalls, dass die verschiedenen Beurteiler unterschiedlich schnell bewerten. Rater 8, der durchschnittlich 8.35 min. für die Bewertung einer Map benötigt, ist als extremer Beurteiler einzuschätzen. Wenn dieser Rater aus den Analysen herausgenommen wird, verringert sich die durchschnittliche Bewertungszeit pro Map von 5.74 min. auf 4.90 min.

Der Beurteilungsbogen bietet eine Möglichkeit, Concept Maps zeitökonomisch effektiv zu beurteilen. Weitere Ergebnisse, die die zeitliche Wirtschaftlichkeit unterstützen, werden für die Studie 2 auf Seite 88 diskutiert.

Computergestützte Auswertung von Concept Maps

Die in dieser Studie eingesetzte Software AKOVIA ermöglicht die Berechnung von semantischen Parametern, die ein Maß für die inhaltliche Qualität von Concept Maps sein sollen (vgl. Abschnitt 4.1.3 Computergestützte Auswertung von Concept Maps). Die Ergebnisse zeigen, dass sich die berechneten Maße nicht eignen, um empirisch haltbare Aussagen zur Validität zwischen Beurteilungsbogen, Kompetenztest und den semantischen Parametern treffen zu können. Der Einsatz in dieser Studie hat gezeigt, dass die PC-generierten Parameter keinen Hinweis auf eine inhaltliche Qualität von Concept Maps geben können, sie wurden in dieser Studie deshalb nicht eingesetzt.

Je nach Untersuchungsziel, kann AKOVIA eine Alternative in der Auswertung von Concept Maps sein. Beispielsweise ist ein Einsatz denkbar, wenn z. B. die Anzahl der Begriffe mit einer Modalmap verglichen werden sollen.

Studie 2

Zeitökonomie und Normalverteilung

18 von 39 Physiklehrkräften der Gruppen 1, 2 und 3 melden die Dauer ihrer jeweiligen Schülerbeurteilung zurück. Lehrkräfte, die Concept Maps mittels Beurteilungsbogen bewerten sollen (Gruppe 1), benötigen im Durchschnitt 3.81 min. pro Schülerin und Schüler. Dies ist eine deutliche Zeiteinsparung im Verhältnis zu den studentischen Mitarbeitern, die in Studie 1 5.74 min. brauchten. Die zeitliche Verbesserung lässt sich ebenfalls für die Gruppen 2 (3.38 min./Map) und 3 (3.30 min./Bogen) zeigen, die nur eines der Instrumente nutzen. Die Befunde lassen darauf schließen, dass mit dem Beurteilungsbogen eine schnelle und systematische Beurteilung von Schülerinnen und Schülern ohne vorherige zeitintensive Schulung in der Nutzung von Concept Maps und dem Beurteilungsbogen möglich ist.

Die Leistung der Schülerstichprobe im Kompetenztest und im KFT wird mittels zweier Verfahren auf Normalverteilung untersucht. Der Kolmogorov-Smirnov-Test und die graphischen Betrachtungen zeigen, dass es sich bei beiden Instrumenten um keine normalverteilte Schülerleistung handelt. Dies kann mit dem Erhebungszeitpunkt, wann die Schülerinnen und Schüler das Thema Energie im Unterricht behandelt haben, erklärt werden. Die Studie wurde in einem gesamten Schulhalbjahr durchgeführt. Es gibt Klassen, die zu Beginn des Schuljahres getestet wurden. Zu diesem Zeitpunkt hatten die Schülerinnen und Schüler noch keinen Unterricht zum Konzept Energie, sodass sie die entsprechenden Fragen im Test durch Raten lösen. Diejenigen Klassen hingegen, die am Ende des Schuljahres getestet wurden, hatten bereits das Konzept Energie erlernt. Die nicht vorhandene Normalverteilung im KFT kann dadurch erklärt werden, dass viele Schülerinnen und Schüler den Test nicht komplett gelöst haben.

Da in dieser Studie der primär interessierende Fokus auf der Diagnosegenauigkeit der Lehrkräfte liegt, wirkt sich die fehlende Normalverteilung der Schülerleistung nicht darauf aus.

Hypothese 2.1 und Hypothese 2.2

In Hypothese 2.1 wird angenommen, dass die Physiklehrkräfte eine Rangordnung ihrer Schülerinnen und Schüler erstellen können, die der Rangordnung eines Kompetenztests entspricht. Die Lehrerinnen und Lehrer nutzen hierbei die Concept Maps ihrer Schülerinnen und Schüler und den Concept Map-Beurteilungsbogen.

Diese Hypothese kann akzeptiert werden. Physiklehrkräfte, die die Concept Maps ihrer Schülerinnen und Schüler mittels Beurteilungsbogen bewerten (Gruppe 1), weisen generell eine mittlere positive Rangübereinstimmung in Form einer Rangkorrelation zum Kompetenztest auf (= Diagnosegenauigkeit, $Md_{Gruppe\ 1} = .32$). Die Lehrkräfte der anderen Gruppen sind im Median betrachtet ebenfalls in der Lage, angemessene Rangordnungen zu bilden ($Md_{Gruppe\ 2} = .25$, $Md_{Gruppe\ 3} = .43$, $Md_{Gruppe\ 4} = .53$). Da bislang keine Forschungsergebnisse für Lehrerinnen und Lehrer der Physik vorliegen, kann dieses Ergebnis als Ausgangsbasis für weitere Untersuchungen in diesem Bereich dienen.

Als weitere Hypothese (Hypothese 2.2) wird aufgestellt, dass die höchste Rangordnungsübereinstimmung gemessen als Rangkorrelation in der Gruppe 1 (Concept Maps und Beurteilungsbogen) erwartet wird. Dies entspricht der höchsten Diagnosegenauigkeit, gemessen als mittlerer Median der Gruppe 1. Die Ergebnisse zeigen, dass die Rangkorrelation im Median in Gruppe 4 am höchsten ist, in der keines der Instrumente für eine Rangordnungsbildung genutzt wird. Es folgt ein Abfall der Rangkorrelation im Median: Gruppe 3, die nur den Beurteilungsbogen nutzt, weist die zweithöchste Rangkorrelation auf, gefolgt von den Gruppen 1 und 2, die die Concept Maps nutzen. Auf Basis dieser Ergebnisse wird ein generell vorhandener Unterschied in der Diagnosegenauigkeit zwischen den Gruppen sichtbar ($H\ (3) = 10.78$, $p < .05$). Die weitere Betrachtung der Ergebnisse zeigt, dass eine Differenzhöhe zwischen den Gruppen 1 und 4 statistisch nicht begründet werden kann. Beide Gruppen lassen sich in ihrer Diagnosegenauigkeit nicht voneinander trennen. Lediglich die Leistung der Gruppe 2 unterscheidet sich signifikant von der Leistung der Gruppe 4. Die Diagnosegenauigkeit der Gruppe 2 ist deutlich niedriger als die der Gruppe 4.

Aus dem Vergleich zwischen den Gruppen 1 und 4 kann geschlossen werden, dass Physiklehrkräfte bereits diagnostizieren können. Ebenfalls kann gezeigt werden, dass Lehrpersonen, die Concept Maps und Beurteilungsbögen nutzen, eine ähnlich hohe Diagnosegenauigkeit erreichen können wie Lehrpersonen, die keine Instrumente nutzen (keine signifikante Trennung der Gruppen). Hypothese 2.2 kann vor diesem Hintergrund nicht bestätigt werden. Zusätzlich sollte bei der Interpretation der Ergebnisse die Teststärke berücksichtigt werden. Es muss auf Basis der Teststärke ($1\text{-}\beta = .50$) angenommen werden, dass der Test mit einer Wahrscheinlichkeit von mindestens 50% einen Gruppenunterschied gefunden hätte, wenn dieser existiert. Es kann auf dieser Grundlage nicht vollends ausgeschlossen werden, dass sich die Gruppen 1 und 4 in ihrer Diagnosegenauigkeit dennoch unterscheiden. Dies würde bedeuten, dass die Lehrkräfte der Gruppe 4 in ihrer

Diagnosegenauigkeit besser sind, als die Lehrkräfte der Gruppe 1 (dies zeigt sich in Form des höheren Medianwerts der Gruppe 4 im Vergleich zum Medianwert der Gruppe 1). Für die Interpretation der Ergebnisse sollte dies berücksichtigt werden.

Es können verschiedene Gründe diskutiert werden, warum die Diagnosegenauigkeit der Gruppe 1 nicht am höchsten ist:

1. *Der Leistungsunterschied der Diagnosegenauigkeit zwischen den Gruppen ist intuitiv nachvollziehbar.*

Lehrkräfte beurteilen ihre Schülerinnen und Schüler tagtäglich nach Noten und bilden immer wieder Rangfolgen ihrer Schülerinnen und Schüler. Sie nutzen hierzu Vorerfahrungen mit ihrer Schülergruppe oder aktuelle Testergebnisse und mündliche Noten, die im Vorfeld gegeben werden. Dadurch sind Lehrerinnen und Lehrer in der Rangordnungsbildung grundsätzlich erfahren. Die Lehrpersonen der Gruppe 4 nutzen diese Expertise, da sie weder Concept Maps noch den Beurteilungsbogen für die Bildung der Rangfolge ihrer eigenen Schülerinnen und Schüler nutzen können. Es ist nicht unerwartet, dass diese Gruppe eine hohe Diagnosegenauigkeit aufweist. Die Lehrerrangfolge und die Kompetenztestrangfolge passen besser zueinander, da beiden Rangfolgen als Beurteilungsbasis Testleistungen zu Grunde liegen.

Mit der sukzessiven Zunahme eines der angewendeten Instrumente (Concept Maps und Beurteilungsbogen), wird der Unterschied in der Diagnosegenauigkeit in Form des Medians zwischen den Gruppen immer größer. Dies lässt sich dadurch erklären, dass die Lehrkräfte durch die Instrumente ihren Fokus für die Bildung der Rangfolge ändern. Den Lehrkräften der Gruppen 1, 2 und 3 wird durch die Instrumente die Möglichkeit gegeben, sich neben den Testleistungen an weiteren Schülermerkmalen, wie beispielsweise Schülervorstellungen, die in den Concept Maps zu erkennen sind (dies gilt für die Gruppen 1 und 2), zu orientieren. Durch den Beurteilungsbogen sind die Lehrkräfte der Gruppe 3 gezwungen, sich ebenfalls an Schülervorstellungen zu orientieren. Der nachfolgende Abschnitt soll erklären, warum die Unterschiede in den Diagnosegenauigkeiten zwischen bestimmten Gruppen statistisch nicht signifikant werden, obwohl ein genereller Einfluss der Gruppenzugehörigkeit auf die Diagnosegenauigkeit vorhanden ist (vgl. *H*-Test)

2. *Die Stichprobengröße beeinflusst das statistische Ergebnis der Unterschiedsprüfung.*

Die Mediane der Diagnosegenauigkeiten in den Gruppen zeigen Unterschiede. Die statistischen Tests (*U*-Tests) zeigen allerdings nur den Gruppenunterschied zwischen den Gruppen 2 und 4. Die geringe Stichprobengröße von 48 Lehrerinnen und Lehrern, die

ungleich auf die vier Gruppen verteilt sind ($N_{Gruppe\ 1}$ = 13, $N_{Gruppe\ 2}$ = 14, $N_{Gruppe\ 3}$ = 12, $N_{Gruppe\ 4}$ = 9), kann als Grund für die nicht signifikanten Gruppenunterschiede zwischen den Gruppen 3 und 4 angesehen werden. Die Gruppen 1 und 4 weisen einen knappen nicht signifikanten Unterschied in der Diagnosegenauigkeit auf, sodass grundsätzlich davon ausgegangen werden kann, dass eine Erhöhung der Stichprobengröße einen signifikanten Unterschied zwischen diesen Gruppen entstehen lassen könnte. Die Ergebnisse der Teststärken weisen zusätzlich darauf hin, dass die Stichprobengröße der Gruppen zu klein ist, um die gemessenen Effektstärken bei einer erneuten Messung wieder zu messen. Die Stichprobengrößen der Gruppen müssten hierfür vergrößert werden.

Zusätzlich erschweren die nicht-parametrischen Berechnungen die Berechnung der tatsächlichen Stärke der statistischen Tests (vgl. Rasch, Friese, Hofmann & Naumann, 2010) Bereits im Ergebnisteil wird darauf hingewiesen, dass die ermittelten Teststärken nur Mindestteststärken sind. Die Teststärken liegen bei nicht-parametrischen Tests höher, können jedoch nicht zuverlässig berechnet werden, da die für diese Berechnung erforderliche Nutzung des parametrischen *d* zur Berechnung der Teststärke zu einer Unterschätzung der wahren Teststärke in nicht-parametrischen Tests führt. (vgl. Rasch, Friese, Hofmann & Naumann, 2010).

Im Vorfeld wurden durch die Abschätzung der sogenannten Power die Stichprobengröße für die einzelnen Gruppen ermittelt (empfohlene Stichprobengröße: N = 20 Lehrpersonen pro Gruppe). Sie können nicht erreicht werden. Die Lehrerakquise gestaltete sich während der gesamten Studie als schwierig und langwierig. Eine Wiederholung der Studie mit mehr Lehrerinnen und Lehrern könnte weitere Gruppenunterschiede entdecken bzw. bereits nachgewiesene Effekte deutlicher herausarbeiten.

3. *Die niedrigen Rangkorrelationen in den Gruppen 1, 2 und 3 sind darauf zurückzuführen, dass die Lehrkräfte im Umgang mit den Instrumenten ungeübt sind.*

In den Gruppen 1, 2 und 3 sind, zusätzlich zu den niedrigen Medianen, große Spannweiten zwischen den Diagnosegenauigkeiten in den Gruppen festzustellen. Punkt 1 erklärt bereits, dass sich Lehrkräfte bei der Bewertung der Schülerinnen und Schüler an Testleistungen orientieren. Lehrkräfte der Gruppen 1, 2 und 3 können durch die Nutzung von Concept Maps und Beurteilungsbögen zusätzlich ohne Notenorientierung diagnostisch beurteilen, wodurch mehr Möglichkeiten für unterschiedliche Beurteilungen entstehen. Die große Spannweite insbesondere in den Gruppen 1 und 2 kann zusätzlich auf das

ungewohnte Verfahren der Concept Map-Beurteilung zurückgeführt werden. Concept Maps spielen im Physikunterricht bislang keine Rolle.

4. *Die Kenntnis der Schülernamen in den Gruppen 3 und 4 kann eine Wechselwirkung mit der Höhe der Diagnosegenauigkeit erzeugen.*

Die Lehrkräfte der Gruppen 3 und 4 müssen die Namen ihrer Schülerinnen und Schüler kennen, um eine Rangfolge bilden zu können. Um die verschiedenen Untersuchungsgruppen bestmöglich miteinander vergleichen zu können, wurden möglichst viele Faktoren in der Durchführung konstant gehalten. Es ist anzunehmen, dass die Lehrkräfte der Gruppen 3 und 4 durch die Kenntnis der Namen einen Vorteil besitzen, der für die Höhe der Diagnosegenauigkeit nicht unerheblich ist. Ausgehend von diesem Aspekt scheinen in den Gruppen 3 und 4 die erfahrungsbasierten Eindrücke die Rangfolge stärker zu beeinflussen als die instrumentbasierte Einschätzung durch anonymisierte Concept Maps. Dies erklärt neben der ungleichen Stichprobengröße in den Gruppen zusätzlich den nicht vorhandenen Gruppenunterschied zwischen den Gruppen 3 und 4.

Die Nutzung der anonymen Concept Maps bietet jedoch die Möglichkeit, dass Lehrerinnen und Lehrer ihre Schülerinnen und Schüler unbeeinflusst von ihren vorangegangenen Erfahrungen diagnostizieren können. Die im Median mittelhohen Diagnosegenauigkeiten der Gruppen 1 und 2 sind ein Hinweis für erfolgreiches diagnostizieren, das unabhängig von Erfahrung mit den zu beurteilenden Schülerinnen und Schülern erreicht werden konnte. Es ist anzunehmen, dass durch die parallele Nutzung von Concept Maps, dem Beurteilungsbogen und der Erfahrung aus anderen Unterrichtsituationen die dadurch grundsätzlich bereits vorhandene Diagnosegenauigkeit gesteigert und verbessert werden kann (Gruppe 1 im Vergleich mit Gruppe 4).

Um den Effekt einzuschätzen, den die Kenntnis der Namen auf die Höhe der Diagnosegenauigkeit haben kann, müsste zusätzlich zu den bereits bestehenden vier Gruppen, eine fünfte Gruppe untersucht werden. Diese fünfte Gruppe müsste Concept Maps mit den Namen der Schülerinnen und Schüler erhalten und sie mit dem Beurteilungsbogen bewerten. (vgl. Abschnitt 4.2.1 Design der Studie 2).

5. *Der Erhebungszeitpunkt, wann die Lehrkräfte die Schülerinnen und Schüler beurteilen, kann eine Rolle bei der Höhe der Diagnosegenauigkeit spielen.*

Voraussetzung für die Datenerhebung in den Gruppen 3 und 4 ist, dass die Lehrkräfte ihre Schülerinnen und Schüler bereits im Inhaltsbereich Energie unterrichtet haben müssen, um eine themenbezogene Rangordnung der Schülerleistung bilden zu können.

Die Lehrpersonen der Gruppe 3 wurden mehrheitlich im ersten Schulhalbjahr der Einführungsphase nach den Herbstferien besucht. Die Entscheidung diese Gruppe zu diesem Zeitpunkt zu testen, hat mit der schwierigen Lehrerakquise zu tun. Beginn der Studie war das zweite Schulhalbjahr 2011/12, in dem die Mehrheit der Stichprobe getestet wurde. Um jedoch die geplante Größe der Teilstichproben zu erreichen (speziell in Gruppe 3 fehlten noch Lehrerinnen und Lehrer), wurde die Studie in das erste Halbjahr des neuen Schuljahres verlängert. Um den Physiklehrkräften der Gruppe 3 die Möglichkeit zu geben, ihre Schülerinnen und Schüler im Konzept Energie beurteilen zu können, wurden die Lehrpersonen nach den Herbstferien besucht. Allerdings scheint nach der Meinung der Lehrkräfte der Gruppe 3 der Zeitpunkt nach den Ferien für die Lehrkräfte nicht ausreichend gewesen zu sein, um fundiert ihre Schülerinnen und Schüler im Basiskonzept Energie einzuschätzen.

Alle anderen Lehrerinnen und Lehrer wurden, mit wenigen Ausnahmen, im zweiten Schulhalbjahr der Einführungsphase besucht.

6. *Grundsätzlich muss diskutiert werden, ob die Erstellung einer Schülerrangfolge aus dem Kompetenztest mit der Rangfolge vergleichbar ist, die von den Lehrpersonen erzeugt wird.*

In vielen Forschungsarbeiten werden bereits Rangfolgen von Schülertestleistungen mit den Einschätzungen der Lehrkräfte verglichen. Es bleibt fraglich, ob die Rangfolgen jeweils auf Basis des gleichen Konstrukts erstellt werden.

In dieser Studie scheint die Rangfolge der Gruppe 4 an der Rangfolge orientiert zu sein, die ein Kompetenztest erzeugt, während die Lehrpersonen der Gruppen 1, 2 und 3 auf Basis der Concept Maps oder einer Mischung aus Concept Maps und testbasierter Erfahrung und Schülerleistungen in der unmittelbaren Vergangenheit urteilen. Die Vergleichsbasis scheint in den einzelnen Gruppen teilweise unterschiedlich zu sein, weshalb die Rangkorrelationen in diesen Gruppen unterschiedlich hoch ausgeprägt sind.

Aus diesem Grund muss diskutiert werden, ob die entwickelten Instrumente ähnliche Konstrukte messen. In Studie 1 kann gezeigt werden, dass die konvergente Validität zwischen dem Concept Map-Beurteilungsbogen und dem Kompetenztest zufriedenstellend ist. Dies stimmt mit Ergebnissen anderer Forschungsarbeiten zu dieser Thematik überein (vgl. Ruiz-Primo & Shavelson, 1996).

Kontrollvariablen

Die Schülerleistung im Kompetenztest steht teilweise mit den erhobenen Kontrollvariablen in Zusammenhang. Die Leistung im KFT korreliert positiv mittelhoch mit der Schülerleistung im Kompetenztest und den Schulnoten in den Fächern Physik, Mathematik und Deutsch.

Die kognitiven Fähigkeiten in Form des figuralen Denkens korrelieren niedrig mit der in den Concept Maps erbrachten Leistung der Gruppe 1, was darauf hindeutet, dass das figurale Denken nur gering mit der Leistung in den Concept Maps zusammenhängt. Auf Grund des Designs liegen für die anderen Untersuchungsgruppen keine Concept Map-Daten vor (Gruppe 2 erstellt zwar Concept Maps, aber erhält keine Punkte, Gruppen 3 und 4 erstellen keine Concept Maps), sodass der Vergleich nur für die Gruppe 1 durchgeführt werden kann.

Auf Lehrerebene werden Lehrermerkmale wie die Anzahl der Berufsjahre, das Alter, das Geschlecht, die Kenntnis über Concept Maps und die Nutzung von Concept Maps erfasst. Diese Kontrollvariablen stehen in keinem signifikanten Zusammenhang zur Diagnosegenauigkeit. Ebenfalls zeigt sich kein Zusammenhang zwischen der Klassenleistung im Kompetenztest und im KFT mit der Diagnosegenauigkeit einer Lehrperson.

In Übereinstimmung mit der Forschungslage zur Diagnosegenauigkeit leisten die erhobenen Lehrermerkmale keinen Beitrag zur Diagnosegenauigkeit. Bereits bei Schrader (1989) wird erklärt, dass die Lehrermerkmale die Diagnosegenauigkeit nicht beeinflussen. Es lässt sich vielmehr annehmen, dass sich andere Faktoren, wie etwa das fachspezifische Professionswissen oder das fachdidaktische Professionswissen, auf die Diagnosekompetenz und die Diagnosegenauigkeit auswirken können (vgl. u. a. Cappell, 2013, Rath & Reinhold, 2014). Dies sollte in zukünftigen Untersuchungen berücksichtigt werden. Zusätzlich kann die Variable ‚Ausbildungshintergrund' noch keinen Einfluss auf die Diagnosegenauigkeit haben. Die Lehrkräfte, die während ihres Studiums in pädagogischer Diagnostik ausgebildet werden, werden erst noch in den Schuldienst eintreten.

Abschließend kann die Studie 2 als zentrales Ergebnis zeigen, dass Physiklehrkräfte, die Concept Maps und den Beurteilungsbogen für eine Bewertung ihrer Schülerinnen und Schüler nutzen (Gruppen 1, 2, 3), sich nicht in der Diagnosegenauigkeit

von denjenigen Physiklehrerinnen und –lehrern unterscheiden lassen, die erfahrungsbasiert die Bewertung vornehmen (Gruppe 4).

7 Zusammenfassung und Ausblick

Das in dieser Arbeit vorgestellte Projekt wird durch zwei Befunde begründet:

1. Es fehlen derzeit wissenschaftlich erprobte Diagnoseinstrumente für den Physikunterricht, die für die Lehrerinnen und Lehrer konzipiert wurden.

2. Für Lehrkräfte der Fächer Deutsch, Mathematik und Englisch besteht Optimierungsbedarf ihrer Diagnosekompetenz/Diagnosegenauigkeit. Für Physiklehrkräfte können derzeit noch keine Aussagen zur Diagnosegenauigkeit getroffen werden.

Ausgehend von diesen Punkten war das übergeordnete Ziel dieser Studie, ein praxistaugliches Diagnoseinstrument zur Schülerdiagnose für Physiklehrerinnen und -lehrer zu entwickeln. Zusätzlich sollte der Umgang mit den entwickelten Instrumenten über die Diagnosekompetenz der Physiklehrkräfte, gemessen als Diagnosegenauigkeit, evaluiert werden.

Concept Maps werden als ein Instrument angesehen, das für die Diagnose von Schülerkonzepten im laufenden Unterricht geeignet ist. In diesem Projekt wurden Concept Map-Aufgabenformate und ein Instrument entwickelt, mit dem die Lehrerinnen und Lehrer die von ihren Schülerinnen und Schülern generierten Concept Maps bewerten können. In zwei Studien wurde untersucht, inwiefern die in diesem Projekt eingesetzten Concept Map-Aufgabenformate Kompetenzen im Basiskonzept Energie abbilden können (Forschungsfrage 1) und inwiefern Concept Maps ein geeignetes Diagnoseinstrument für Physiklehrerinnen und –lehrer sind (Forschungsfrage 2).

Die Ergebnisse dieses Projekts können teilweise die Forschungsfragen beantworten. Die Entwicklung, Pilotierung und Validierung der Instrumente findet in der ersten Studie statt. Es kann eine konvergente Validität zwischen zwei Concept Map-Aufgabenformaten (A und B) und einem Testinstrument zum Konzept Energie festgestellt werden.

Die zweite Forschungsfrage kann ebenfalls zufriedenstellend beantwortet werden. Die Ergebnisse lassen darauf schließen, dass Physiklehrkräfte bereits vorhandene diagnostische Fähigkeiten besitzen.

Ausgehend von den Ergebnissen lässt sich ein Nutzen von Concept Maps für Lehrkräfte ableiten. Ein alltäglicher Einsatz von Concept Maps in der Schule kann vorgeschlagen werden. Es kann angenommen werden, dass die Lehrkräfte durch die Nutzung der Instrumente weitere Merkmale ihrer Schülerinnen und Schüler bewerten als es die übliche Bewertung für die Notengebung (Tests und Bewertung mündlicher Leistungen) erlaubt. Zusätzlich wird den Lehrpersonen durch den Beurteilungsbogen ein zeitökonomischer Weg für die Beurteilung von Vernetzungsleistungen ermöglicht.

Der Optimierungsbedarf dieser Arbeit liegt im Design der Studie 2. Das Design folgt keinem reinen 2x2-Design. Es fehlen für eine weitere Einschätzung der Ergebnisse mindestens zwei zusätzliche Versuchsgruppen, die den Faktor anonyme/personalisierte Rangordnungsbildung prüfen. Zusätzlich sind die Stichprobengrößen zu optimieren. Ebenfalls müssen weitere Auswerteverfahren der Concept Maps erprobt werden (beispielsweise eine graphentheoretische Auswertung, die den Inhalt einer Concept Map abbilden kann) und weitere Validierungsinstrumente, um eine bessere Passung zwischen dem Concept Map-Aufgabeformat und Bewertungsformat und einem externen Instrument zu erzielen.

Das Projekt zeigt, dass Concept Maps weiterhin schwer zu beurteilen sind, jedoch mit dem in diesem Projekt entwickelten papierbasierten Bewertungsverfahren angemessen bewertet werden können. Wünschenswert wäre, die Concept Maps mittels ‚schnellem Mausklick' vollautomatisiert durch einen PC auswerten zu können. Die derzeit zur Verfügung stehenden Computerprogramme können die inhaltliche Qualität der Concept Maps nicht abbilden. Die in dieser Untersuchung parallel eingesetzte, sich in der Weiterentwicklung befindlichen Software AKOVIA (Ifenthaler, 2010) konnte die versprochenen Ergebnisse nicht erbringen.

Die Ergebnisse zur Diagnosegenauigkeit können als Ausgangspunkt für weitere Forschung genutzt werden. Es stellt sich die Frage, wie Physiklehrerinnen und –lehrer ihren Unterricht nach der Diagnose adaptieren. Eine gezielte Förderung von Schülerinnen und Schülern setzt eine erfolgreiche Diagnose voraus. Die Lehrkräfte, die Concept Maps zur Diagnose einsetzen, können in weiteren Studien beispielsweise bezüglich ihrer Fähigkeiten zum angemessenen Adaptieren ihres Unterrichts untersucht werden.

Der Befund dieser Arbeit, dass Physiklehrkräfte bereits über eine Diagnosegenauigkeit verfügen, ist erfreulich. Es ist allerdings offen, wie die Diagnosegenauigkeit der Lehrkräfte im Bereich der Niveaukomponente und der Streuungskomponente ausgeprägt ist. Ebenfalls sollte eine detailliertere Modellierung der Diagnosekompetenz verfolgt werden. Die von Schrader und Helmke (1987) vorgeschlagene Modellierung der Diagnosekompetenz in Form der Diagnosegenauigkeit ist nur begrenzt hilfreich, sie kann das Konstrukt ‚Diagnosekompetenz' nur teilweise charakterisieren. Hinsichtlich der Lehreraus- und –fortbildung sollten Programme entwickelt werden, die die bereits vorhandene Diagnosegenauigkeit weiter fördern und optimieren können, speziell für Lehrpersonen, die in den Beruf eingestiegen sind oder für

Studierende. Hierzu sollten ebenfalls die in den Standards zur Lehrerbildung der KMK (Kompetenzbereich 7 ‚Beurteilen', 2004) formulierten Kompetenzen auf ihre Umsetzung und Umsetzbarkeit überprüft werden.

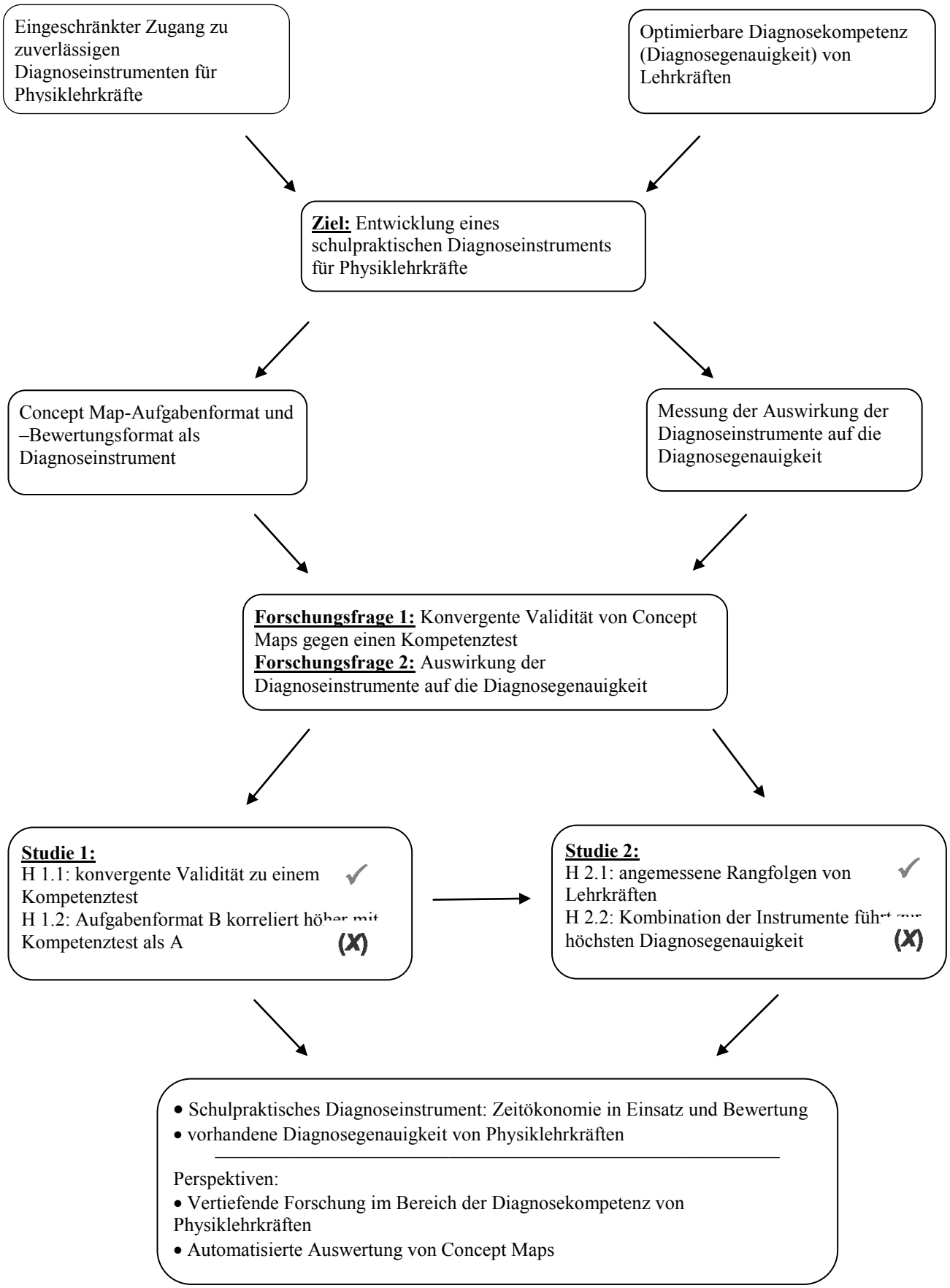


Abbildung 7.1. Zusammenfassende Darstellung der wesentlichen Elemente dieser Arbeit.

8 Abbildungsverzeichnis

9 Tabellenverzeichnis

10 Literaturverzeichnis

Abs, H. J. (2007). Überlegungen zur Modellierung diagnostischer Kompetenz bei Lehrerinnen und Lehrern. In: M. Lüders, J. Wissinger (Hrsg.): *Forschung zur Lehrerbildung. Kompetenzentwicklung und Programmevaluation* (S. 63-84). Waxmann: Münster.

Acton, W., Johnson, P. & Goldsmith, T. (1994). Structural Knowledge Assessment. In: *Journal of Educational Psychology 86* (2). 303-311.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention*. Berlin: Springer.

Anders, Y., Kunter, M., Brunner, M., Krauss, S. & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. In: *Psychologie in Erziehung und Unterricht 57*. 175-193.

Anderson, T. H. & Huang, S.-C. C. (1989). *On using Concept Maps to assess the Comprehension Effects of Reading Expository Text* (Technical report No. 483). Urbana-Champaign: Center for the studying of reading, University of Illinois at Urbana-Champaign. (ERIC Document Reproduction Service No. ED 310 368).

Artelt, C. & Gräsel, C. (2009). Diagnostische Kompetenz von Lehrkräften. In: *Zeitschrift für Pädagogische Psychologie 23* (3-4). 157-160.

Ausubel, D. P. (1960). The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material. In: *Journal of Educational Psychology 51* (5). 267-272.

Baddeley, A. (1992). Working Memory. In: *Science 31* (255). 556-559.

Barenholz, H. & Tamir, P. (1992). A comprehensive use of Concept Mapping in Design Instruction and Assessment. In: *Research in Science & Technological Education 10* (1). 37-52.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.) (2001). *Deutsches PISA-Konsortium. PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske + Budrich.

Becker, G., Horstkemper, M., Risse, E., Stäudel, L., Werning, R. & Winter, F. (2006). *Diagnostizieren und Fördern. Stärken entwickeln-Können entwickeln.* Seelze: Erhard Friedrich GmbH.

Behrendt, H. & Reiska, P. (2001). Abwechslung im Naturwissenschaftsunterricht mit Concept Mapping. In: *PLUS LUCIS 1*. 9-12.

Berkemeyer, N., Bos, W., Holtappels, H. G., Meetz, F. & Rollett, W. (2010). „Ganz In“: Das Ganztagsgymnasium in Nordrhein-Westfalen-Bestandsaufnahme und Perspektiven eines Schulentwicklungsprojekts. In: N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvany, R. Schulz-Zander (Hrsg.). *Jahrbuch der Schulentwicklung. Band 16* (S. 131-153). Weinheim: Juventa Verlag.

Beyerbach, B. A. & Smith, J. M. (1990). Using a Computerized Concept Mapping Program to assess Preservice Teachers‘ Thinking about Affective Teaching. In: *Journal of Research in Science Teaching 27* (10). 961-971.

Bonato, M. (1990). *Wissensstrukturierung mittels Struktur-Lege-Techniken. Eine graphentheoretische Analyse von Wissensnetzen.* Frankfurt am Main: Peter Lang GmbH.

Borgatti, S. P. & Everett, M. G. (2006). A Graph-Theoretic Perspective on Centrality. In: *Social Networks 28.* 466-484.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation.* Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler.* Heidelberg: Springer.

Bremm, M. H. & Kühn, R. (1992). *Rechentest RT 9+.* Weinheim: Beltz.

Brennan, M. M. & Redding, K. R. (1985). *Are Teachers good Predictors of School Level or Statewide Level of Student Performance?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Bühl, A. (2010). *PASW 18. Einführung in die moderne Datenanalyse.* München: Pearson Studium.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson.

Bühner, M. & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler.* München Pearson.

Cappell, J. (2013). *Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase.* Berlin: Logos.

Cappell, J. & von Aufschnaiter, C. (2011). Diagnosekompetenz angehender Physiklehrkräfte. In: D. Höttecke (Hrsg.). *Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie* (S. 78-80). Berlin: LIT Verlag.

Cappell, J. & von Aufschnaiter, C. (2012). Die Entwicklung diagnostischer Kompetenz von angehenden Physiklehrer/innen. In: S. Bernholt (Hrsg.). *Konzepte fachdidaktischer Strukturierung für den Unterricht* (S. 239-241). Berlin: LIT Verlag.

Cohen, J. (1988). *Statistical Power for the Behavioral Sciences* (2 ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Conradty, C. & Bogner, F. X. (2012). Knowledge presented in Concept Maps: Correlations with conventional Cognitive Knowledge Tests. In: *Educational Studies 38* (3). 341-354.

Creswell, J. W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research.* Upper Saddle River, NJ: Pearson.

Eckert, A. (2000). Die Netzwerk-Elaborierungs-Technik (NET)-Ein computerunterstütztes Verfahren zur Diagnose komplexer Wissensstrukturen. In: H. Mandl, F. Fischer (Hrsg.). *Wissen sichtbar machen* (S. 137-157). Göttingen: Hogrefe Verlag.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. In: *Behavior research Methods 39* (2). 175-191.

Field, A. (2009). *Discovering Statistics Using SPSS.* London: SAGE Publications.

Fischler, H. & Peuckert, J. (2000). Concept Mapping in Forschungszusammenhängen. In: H. Fischler, J. Peuckert (Hrsg.). *Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie* (S. 1-21).Berlin: Logos.

Fisher, K. M. (1990)**.** Semantic Networking: the New Kid in the Block. In: *Journal of Research in Science Teaching 27* (10). 1001-1018.

Fraenkel, J. R., Wallen, N. E. & Hyun, H. N. (2012). *How to design and evaluate Research in Education.* New York: McGraw-Hill.

Friege, G. & Lind, G. (2000). Begriffsnetze und Expertise. In: H. Fischler, J. Peuckert (Hrsg.). *Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie* (S. 147-178).Berlin: Logos.

Fürstenau, B. & Trojahner, I. (2005). Prototypische Netzwerke als Ergebnis struktureller Inhaltsanalysen. In: P. Gonon, F. Klauser, R. Nickolaus, R. Huisinga (Hrsg.): *Kompetenz, Kognition und neue Konzepte der beruflichen Bildung* (S. 191-202). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gläser-Zikuda, M. (2010). Lernvoraussetzungen diagnostizieren und Fördermaßnahmen realisieren. In: T. Bohl, W. Helsper, H. G. Holtappels, C. Schelle (Hrsg.). *Handbuch Schulentwicklung. Theorie-Forschungsbefunde-Entwicklungsprozesse-Methodenrepertoire* (S. 369-376). Bad Heilbrunn: Klinkhardt.

Greve, W. & Ventura, D. (1995). *Wissenschaftliche Beobachtungen. Eine Einführung.* Weinheim: Psychologie Verlags Union.

Handcock, M. S., Hunter, D. R., Butts, C., Goodreau, S. M. & Morris, M. (2008). statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. In: *Journal of Statistical Software 21*(1). 1-11.

Haschke-Hirth, A. & Kuhle, C. (2010). Diagnostische Kompetenzen. Unterricht-Diagnose-Kompetenz (UDiKom). KMK-Projekt zur Stärkung diagnostischer Kompetenzen von Lehrkräften. In: *Schule NRW 04/10.* 173-174.

Hattie, J. (2009*). Visible Learning. A Synthesis of over 800 Meta-Analyses relating to Achievement.* New York: Routledge.

Haugwitz, M. (2009*). Kontextorientiertes Lernen und Concept Mapping im Fach Biologie.* Zugriff auf http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate- 23401/Dissertation_Haugwitz.pdf (01.03.2013).

Haugwitz, M. & Sandmann, A. (2009). Kooperatives Concept Mapping in Biologie: Effekte auf den Wissenserwerb und die Behaltensleistung. In: *Zeitschrift für Didaktik der Naturwissenschaften 15*. 89-107.

Heller, K. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen. KFT 4-12+R.* Göttingen: Open University Press.

Helmke, A. (2009a). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze- Velber: Klett Kallmeyer.

Helmke, A. (2009b). Diagnosekompetenz von Lehrern. In: *PROFIL März 2009.* 32-38.

Helmke, A. (2009c). Die pädagogische Diagnostik führt ein Schattendasein. In: *Frankfurter Allgemeine Zeitung Januar 2009* (6). 8-9.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In: R. Arnold (Hrsg.): *Schulleitung und Schulentwicklung/ Voraussetzungen, Bedingungen, Erfahrungen* (S. 119-144). Hohengehren: Schneider.

Hesse, I. & Latzko, B. (2009). *Diagnostik für Lehrkräfte.* Opladen & Farmington Hills: Barbara Budrich.

Hoge, R. D. (1983). Psychometric Properties of Teacher-Judgement Measures of Pupil Aptitudes, Classroom Behaviors, and Achievement Levels. In: *Journal of Special Education 17* (4). 401-429.

Hoge, R. D. & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. In: *Review of Educational Research 59* (3). 297-313.

Holtappels, H. G. (2004). Deutschland auf dem Weg zur Ganztagsschule?. In: *PÄDAGOGIK 2*. 6-10.

Hopkins, K. D., George, C. A. & Williams, D. D. (1985). The Concurrent Validity of standardized Achievement Tests by Content Area using Teachers' Ratings as Criteria. In: *Journal of Educational Measurement 22* (3). 177-182.

Horton, P. B., McConney, A. A., Gallo, M., Woods, A. L., Senn, G. J. & Hamelin, D. (1993). An Investigation of the Effectiveness of Concept Mapping as an Instructional Tool. In: *Science Education 77* (1). 95-111.

Hucke, L. & Fischer, H. E. (2003). The link of theory and practice in traditional and in computer-based university laboratory experiments. In: D. Psillos, H. Niedderer (eds.). *Teaching and Learning in the Science Laboratory* (S. 205-218). Dordrecht: Kluwer Academic Publishers.

Hucke, L. & Fischer, H. E. (2000). Wissenserwerb und Handlungsregulation im physikalischen Praktikum. In: H. Fischler, J. Peuckert (Hrsg.). *Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie* (S. 57-90). Berlin: Logos.

IBM (2010). IBM Statistics SPSS 18 [Computer Software].

IBM (2012). IBM Statistics SPSS 21 [Computer Software].

Ifenthaler, D. (2010). Relational, Structural, and Semantic Analysis of Graphical Representations and Concept Maps. In: *Educational Technology Research and Development 58* (1). 81-97.

Ingeç, S. K. (2009). Analysing Concept Maps as an Assessment Tool in Teaching Physics and Comparison with the Achievement Tests. In: *International Journal of Science Education 31* (14). 1897-1915.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik.* Weinheim und Basel: Beltz Verlag.

Institute for Human & Machine Cognition (2010). CMap Tools v5.04.01 [Computer Software].

Jäger, R. S. (2009). Diagnostische Aufgaben und Kompetenzen von Lehrkräften. In: K.-H. Arnold, U. Sandfuchs, J. Wiechmann (Hrsg.). *Handbuch Unterricht* (S. 471-476). Bad Heilbrunn: Verlag Julius Klinkhardt.

Jäger, R. S. & Petermann, F. (1995). *Psychologische Diagnostik.* Weinheim: Beltz Verlag.

Jüngst, K. L. & Strittmatter, P. (1995). Wissensstrukturdarstellung: Theoretische Ansätze und praktische Relevanz. In: *Unterrichtswissenschaft 23* (3). 194-207.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. In: *Zeitschrift für Pädagogische Psychologie 19* (1/2). 197-209.

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrkräften.* Münster: Waxmann.

Kauertz, A. & Fischer, H. E.(2010). Standards und Physikaufgaben. In: E. Kircher, R. Girwidz, P. Häußler. (Hrsg.). *Physikdidaktik. Theorie und Praxis.* (S. 663-688). Heidelberg: Springer.

KMK. Sekretariat der Ständigen Konferenz der Kulturminister der Länder in der Bundesrepublik Deutschland (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004.* Bonn.

Kliemann, S. (2008). *Diagnostizieren und Fördern in der Sekundarstufe I*. Berlin: Cornelsen Verlag.

Komorek, M. & Michaelis, J. (2011). Verbundprojekt OLAW-Entwicklung von Diagnose- und Förderkompetenz. In: D. Höttecke (Hrsg.). *Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie* (S. 211-213). Berlin: LIT Verlag.

Langfeldt, H.-P. (2006). *Psychologie für die Schule*. Weinheim: Beltz Verlag.

Langfeldt, H.-P. & Trolldiener, H.-P. (1993).*Pädagogisch-psychologische Diagnostik. Aktuelle Entwicklungen und Ergebnisse.* Heidelberg: Asanger Roland Verlag.

Lay-Dopyera, M. & Beyerbach, B. (1983). *Concept Mapping for individual Assessment.* Syracusem NY: School of Education, Syracuse University. (ERIC Document Reproduction Service No. ED 229 399).

Leutner, D. (2001). Pädagogisch-psychologische Diagnostik. In: D. H. Rost (Hrsg.). *Handwörterbuch Pädagogische Psychologie* (S. 521-530). Weinheim: Verlagsgruppe Beltz.

Liu, X. & McKeough, A. (2005). Developmental Growth in Students‘ Concept of Energy: Analysis from selected Items from the TIMSS Database. In: *Journal of Research in Science Teaching 45* (5). 493-517.

Lomask, M., Baron, J. B., Greig, J. & Harrison, C. (1992). *ConnMap: Conneticut's use of Concept Mapping to assess the Structure of Students' Knowledge of Science.* Paper presented at the annual meeting of the National Association of Research in Science Teaching.Cambridge, MA.

Lukesch, H. (1994). *Einführung in die pädagogisch-psychologische Diagnostik.* Regensburg: CH-Verlag.

Mandl, H. & Fischer, F. (Hrsg.) (2000). *Wissen sichtbar machen*. Göttingen: Hogrefe Verlag.

Markham, K. M., Mintzes, J. J. & Jones, M. G. (1994). The Concept Map as a Research and Evaluation Tool: Further Evidence of Validity. In: *Journal of Research in Science Teaching 31 (1)*. 91-101.

Mavanga, G. G. (2001). *Entwicklung und Evaluation eines experimentell- und phänomenorientierten Optikcurriculums.* Berlin: Logos.

May, P. (2007). *HSP 5-9. Hamburger Schreib-Probe zur Erfassung der grundlegenden Rechtschreibstrategien.* Seelze: vpm.

Mayer, R. & Moreno, R. (2003). Nine Ways to reduce Cognitive Load in Multimedia Learning. In: *Educational Psychologist 38* (1). 43-52.

McClure, J. R. & Bell, P. E. (1990). *Effects of an Environmental Education-related STS Approach Instruction on cognitive Structures of Preservice Science Teachers.* University Park, PA: Pennsylvania State University. (ERIC Document Reproduction Service No. ED 341 582).

McClure, J. R., Sonak, B. & Suen, H. K. (1999). Concept Map Assessment of Classroom Learning: Reliability, Validity and Logistical Practicality. In: *Journal of Research in Science Teaching 36* (4), 475-492.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H. & Ulrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften. In: *Zeitschrift für Pädagogische Psychologie 19* (1/2). 223-235.

Neumann, K., Viering, T. & Fischer, H. E. (2010). Die Entwicklung physikalischer Kompetenz am Beispiel des Energiekonzepts. In: *Zeitschrift für die Didaktik der Naturwissenschaften 16.* 285-298.

Neumann, K., Viering, T., Boone, W. J. & Fischer, H. E. (2013). Towards a Learning Progression in Energy. In: *Journal of Research in Science Teaching 50* (2). 162-188.

Nesbit, J. C. & Adesope, O. O. (2006). Learning with Concept Maps and Knowledge Maps: A Meta-Analysis. In: *Review of Educational Research 76* (3). 413-448.

Novak, J. D. (1990). Concept Mapping: a useful Tool for Science Education. In: *Journal of Research in Science Teaching 27* (10). 937-949.

Novak, J. D. & Gowin, D. B. (1984). *Learning how to learn.* Cambridge: Cambridge University Press.

Novak, J. D., Gowin, D. B. & Johansen, G. T. (1983). The Use of Concept Mapping and Knowledge Vee Mapping with Junior High School Science Students. In: *Science Education 67* (5). 625-645.

Nückles, M., Gurlitt, J., Pabst, T. & Renkl, A. (2004). *Mind Maps und Concept Maps. Visualisieren-Organisieren-Kommunizieren*. München: dtv Verlag, Beck.

Paas, F., Tuovinen, J., Tabbers, H., Van Gerven & P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. In: *Educational Psychologist 38* (1). 63-71.

Paradies, L., Linser, H. J. & Greving, J. (2009). *Diagnostizieren, Fordern und Fördern.* Berlin: Cornelsen Verlag.

Patterson, M. E., Dansereau, D. F. & Newbern, D. (1992). Effects of Communication Aids and Strategies on Cooperative Teaching. In: *Journal of Educational Psychology 84.* 453-461.

Peuckert, J. (1999). Concept Mapping-Lernen wir unsere Schüler kennen!. In: *Physik in der Schule 37* (1). 47-128.

Plötzner, R., Leuders, T. & Wichert, A. (Hrsg.). (2009). *Lernchance Computer-Strategien für das Lernen mit digitalen Medienverbünden*. Münster: Waxmann Verlag.

Pospeschill, M. & Spinath, F. M. (2009): *Psychologische Diagnostik.* München: Ernst Reinhardt.

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2010). *Quantitative Methoden. Band 2. Einführung in die Statistik für Psychologen und Sozialwissenschaftler.* Heidelberg: Springer.

Rath, V. & Reinhold, P. (2014). Diagnosekompetenz von Physiklehramtsstudierenden. In: S. Bernholt (Hrsg.). *Naturwissenschaftliche Bildung zwischen Science- und*

Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung München 2013. (S. 441-443). IPN: Kiel.

Renkl, A. & Nückles, M. (2006). Lernstrategien der externen Visualisierung. In: H. Mandl, H. F. Friedrich (Hrsg.). *Handbuch Lernstrategien* (S. 135-150). Göttingen: Hogrefe Verlag.

Rice, D. C., Ryan, J. M. & Samson, S. M. (1998). Using Concept Maps to Assess Student Learning in the Science Classroom: Must different Methods compete?. In: *Journal of Research in Science Teaching 35* (10). 1103-1127.

Ruiz-Primo, M. (2000). On the use of Concept Maps as an Assessment Tool in Science. What we have learned so far. In: *Revista Electronica de Investigacion Educativa 2* (1). Zugriff auf http://redie.uabc.mx/vol2no1/contents-ruizpri.html. (01.03.2013).

Ruiz- Primo, M. A. & Shavelson, R. J. (1996). Problems and Issues in the Use of Concept Maps in Science Assessment. In: *Journal of Research in Science Teaching 33* (6). 569-600.

Ruiz- Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the Reliability and Validity of Scores from two Concept-Mapping Techniques. In: *Journal of Research in Science Teaching 38* (2). 260-278.

Rost, D. H. (2005). *Interpretation und Bewertung pädagogisch-psychologischer Studien.* Weinheim: Beltz.

Schadé, J. P. (2002). *Lexikon Medizin und Gesundheit : Erste Hilfe, Krankheiten: Ursachen und Behandlungen, Anatomie des Menschen, Wirkstoffe, Arzneimittel, Behandlungsmethoden.* Köln: Serges Medien GmbH.

Schau, C. & Mattern, N. (1997). Use of Map Techniques in Teaching Applied Statistics Courses. *The American Statistician 51* (2). 171-175. Zugriff auf http://dx.doi.org/10.1080/00031305.1997.10473955. (22.02.2013).

Schecker, H. & Klieme, E. (2000). Erfassung physikalischer Kompetenz durch Concept-Mapping-Verfahren. In: H. Fischler, J. Peuckert (Hrsg.). *Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie* (S. 23-56). Berlin: Logos.

Scheele, B. & Groeben, N. (1984). *Die Heidelberger Struktur-Lege-Technik (SLT).* Weinheim: Beltz.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik.* Berlin: Springer Verlag.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts.* Frankfurt am Main: Verlag Peter Lang GmbH.

Schrader, F.-W. (2001). Diagnostische Kompetenz von Eltern und Lehrern. In: D. H. Rost. (Hrsg.). *Handwörterbuch Pädagogische Psychologie* (S. 91-96). Weinheim: Beltz.

Schrader F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. In: *Empirische Pädagogik 1* (1). 27-52.

Segerer, R., Marx, A. & Marx, P. (2012). Unlösbare Items im KFT 4-12+R. In: *Diagnostica 58* (1). 45–50.

Sedlmeier, P. & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson Studium.

Seidel, T. & Prenzel, M. (2007). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen- Erfassung pädagogisch- psychologischer Kompetenzen mit Videosequenzen. In: *Zeitschrift für Erziehungswissenschaft 10.* Sonderheft 8. 201-216.

Siemes, A. (2008). Diagnosetheorien. In: S. Kliemann (2008). *Diagnostizieren und Fördern in der Sekundarstufe I* (S. 12-21). Berlin: Cornelsen Verlag.

Starr, M. L. & Krajik, J. S. (1990). Concept Maps as a Heuristic for Science Curriculum Development: Toward Improvement in Process and Product. In: *Journal of Research in Science Teaching 27* (10). 987-1000.

Stracke, I. (2004). *Einsatz computerbasierter Concept Maps zur Wissensdiagnose in der Chemie. Empirische Untersuchungen am Beispiel des Chemischen Gleichgewichts.* Münster: Waxmann.

Sumfleth, E. & Tiemann, R. (2000). Own Word Mapping- ein alternativer Zugang zu Schülervorstellungen. In: H. Fischler, J. Peuckert (Hrsg.). *Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie* (S. 179-204). Berlin: Logos.

Sumfleth, E., Neuroth, J. & Leutner, D. (2010). Concept Mapping-eine Lernstrategie muss man lernen. In: *CHEMKON 17* (2), 66-77.

Südkamp, A., Möller, J. & Pohlmann, B. (2008). Der Simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz. In: *Zeitschrift für Pädagogische Psychologie 19* (1/2). 261-276.

Tent, L. & Stelzl, I. (1993). *Pädagogisch- psychologische Diagnostik. Band 1. Theoretische und methodische Grundlagen.* Göttingen: Hogrefe.

Tergan, S.-O. (2006). Individuelles Wissens- und Informationsmanagement mit Concept Maps bei ressourcenbasierten Lernen. In: H. Mandl, H. F. Friedrich (Hrsg.). *Handbuch Lernstrategien* (S. 307-324). Göttingen: Hogrefe Verlag.

Tiemann, R. (1999). *Analyse individueller Wissensstrukturen im Kontext Chemie mit Hilfe eines neuen Mapping-Verfahrens.* Münster: lit.

Trochim, W. M. K. (1989). An Introduction to Concept Mapping for planning and evaluation. In: *Evaluation and Program Planning 12.* 1-16.

Viering, T. (2012). *Entwicklung physikalischer Kompetenz in der Sekundarstufe I. Validierung eines Kompetenzentwicklungsmodells für das Energiekonzept im Bereich Fachwissen.* Berlin: Logos.

Wahser, I. (2007). *Training von naturwissenschaftlichen Arbeitsweisen zur Unterstützung experimenteller Kleingruppenarbeit im Fach Chemie.* Berlin: Logos.

Walpuski, M., Kauertz, A., Kampa, N., Fischer, H. E., Mayer, J., Sumfleth, E. & Wellnitz, N. (2010). ESNaS- Evaluation der Standards für die Naturwissenschaften in der Sekundarstufe I. In: A. Gehrmann, U. Hericks, M. Lüders (Hrsg.). *Bildungsstandard und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 171-184). Bad Heilbrunn: Julius Klinkhardt.

Weinert, F. E. (1998). Vermittlung von Schülerqualifikationen. In: S. Matalik, D. Schade (Hrsg.). *Entwicklung in Aus- und Weiterbildung: Anforderungen, Ziel, Konzepte* (S. 23-43). Baden-Baden: Nomos.

Weinert, F. E. (2000). Lehren und Lernen für die Zukunft-Ansprüche an das Lernen in der Schule. In: *Pädagogische Nachrichten Rheinland-Pfalz 2*. 1-16.

Weinert, F. & Schrader, F. (1986). Diagnose des Lehrers als Diagnostiker. In: H. Petillon, J. W. Wagner, B. Wolf (Hrsg.). *Schülergerechte Diagnose* (S. 11-29). Theoretische und empirische Beiträge zur Pädagogischen Diagnostik. Weinheim: Beltz Verlag.

Weir, J. P. (2005). Quantifying Test-Retest Reliability using the Intraclass Correlation Coefficient and the SEM. In: *Journal of Strength and Conditioning Research 19* (1). 231-240.

Wild, K.-P. & Krapp, A. (2006). Pädagogisch-psychologische Diagnostik. In: A. Krapp, B. Weidemann (Hrsg.). *Pädagogische Psychologie* (S. 525-574). Weinheim: Beltz Verlag.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

11 Anhang

A. Instrumente

A1. Concept Map-Aufgabenformat

A2. Concept Map-Beurteilungsbogen

A3. Lehrerfragebogen zu Ausbildung und Beruf

A4. Manual zur Nutzung des Concept Map-Beurteilungsbogens

A5. Rankingbögen der verschiedenen Gruppen

B. Ergebnisse

B1. Studie 1-nicht-parametrische Berechnungen

B2. Studie 2-parametrische Berechnungen

A. Instrumente

Es folgen die Instrumente und Materialien, die für die Physiklehrerinnen und -lehrer in diesem Projekt entwickelt wurden.

A.1 Concept Map-Aufgabenformat

Aufgabenformat A

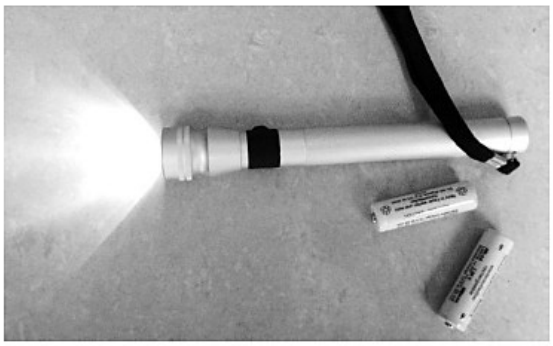
Liebe Schülerin, lieber Schüler, nur Liste, 10 aus 21

dies sind die Begriffe, mit denen du die Concept Map erstellen sollst. Zunächst suche dir aus dieser Liste ***nur 10 Begriffe*** aus, mit der du Map erstellst (unterstreiche sie unten in der Liste).
Wenn du mit der Map fertig bist, kannst du freiwillig noch weitere Begriffe, die dir einfallen, einbauen. Du kannst auch die übrigen Begriffe aus der Liste nehmen. Nutze aber zunächst nur die 10 Begriffe, die du dir ausgesucht hast!

Vielen Dank!

-Energieform/Energieart
-Energiespeicher
-Bewegungsenergie
-Lageenergie
-Motor
-Temperatur
-Leistung
-Thermische Energie
-Energieträger
-Körper/Gegenstand
-Energiewandler
-Generator
-Physikalische Arbeit
-Geschwindigkeit
-Kraft
-Energieverlust
-System
-Strahlungsenergie
-Chemische Energie
-Höhe
-Batterie
-Treibstoff
-Licht
-Elektrische Energie
-Nahrung
-Glühbirne

Aufgabenformat B
-PHASE 1-



Liebe Schülerin, lieber Schüler,

dies ist deine Aufgabe:

Betrachte die Bilder unter dem Aspekt Energie. Erstelle dazu dann eine Concept Map zum Begriffsfeld Energie.

-PHASE 2-

Liebe Schülerin, lieber Schüler, erst Assoziation, dann Liste, 10 aus 21

jetzt wechselst du bitte deine Stiftfarbe. Dies sind die Begriffe, mit denen du deine Concept Map erweitern sollst.
1. Schau, welche Begriffe aus der Liste du bereits in deiner Concept Map hast.
2. Ordne diese Begriffe den Begriffen in deiner Concept Map zu, z.B. schreibe sie als eigenständigen Begriff dran.
3. Dann ergänze deine Concept Map um **mindestens 10** weitere Begriffe aus der Liste. (*Kreise* alle Begriffe in der Map ein).
Wenn du mit der Map fertig bist, kannst du freiwillig noch weitere Begriffe, die dir einfallen, einbauen. Du kannst auch die übrigen Begriffe aus der Liste nehmen. Nutze aber zunächst nur die 10 Begriffe, die du dir ausgesucht hast!

Vielen Dank!

-Energieform/Energieart
-Energiespeicher
-Bewegungsenergie
-Lageenergie
-Motor
-Temperatur
-Leistung
-Thermische Energie
-Energieträger
-Körper/Gegenstand
-Energiewandler
-Generator
-Physikalische Arbeit
-Geschwindigkeit
-Kraft
-Energieverlust
-System
-Strahlungsenergie
-Chemische Energie
-Höhe
-Batterie
-Treibstoff
-Licht
-Elektrische Energie
-Nahrung
-Glühbirne

A.2 Concept Map-Beurteilungsbogen

Studie 1 (lange Version)

Teilnehmernummer:_____ MapID:_____Ratingsheet- Gesamtkonzept„Energie_L"

Liebe Lehrerin, lieber Lehrer.

Im Folgenden finden Sie Aussagen, die Sie bitte auf Grundlage der Concept Map Ihrer Schülerin bzw. Ihres Schülers bewerten. Bitte kreuzen Sie jede Aussage an, inwiefern sie für die Concept Map der Schülerin bzw. der Schülers zutrifft oder nicht.

Die Schülerin/ Der Schüler ...

	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
1. ...hat erkannt, dass Objekte bzw. Körper über Energie verfügen können, z. B. ein Vogel sitzend auf einem Ast besitzt Lageenergie.	☐	☐	☐	☐
2. ...kann verschiedene Energiequellen benennen, z. B. Sonne als Energiequelle, Nahrung als Energiequelle.	☐	☐	☐	☐
3. ...hat erkannt, dass die Energie von Objekten bzw. Körpern in verschiedenen Formen auftreten kann, z. B. eine Achterbahn mit Looping besitzt zu verschiedenen Zeiten Lage- &Bewegungsenergie.	☐	☐	☐	☐
4. ...kann verschiedene Energieformen benennen, z. B. thermische Energie, elektrische Energie, Lageenergie, Bewegungsenergie.	☐	☐	☐	☐
5. ...hat erkannt, dass die Energie der Objekte in Verbindung mit bestimmten Eigenschaften des Objektes bzw. Körpers steht, z. B. ruht ein Stein besitzt er Lageenergie, rollt er, besitzt er Bewegungsenergie.	☐	☐	☐	☐
6. ...kann verschiedene Eigenschaften eines Objektes den verschiedenen Energieformen in einer gegebenen Situation zuordnen, z. B. ein gespannter Bogen besitzt Spannenergie.	☐	☐	☐	☐
7. ...hat erkannt, dass Energie notwendig ist, um Arbeit zu verrichten, z. B.Energie ist die Fähigkeit, um Arbeit zu verrichten.	☐	☐	☐	☐
8. ...hat erkannt, dass Energie in Energieträgern transportiert wird, z. B. Nahrung oder Kohle sind Energieträger.	☐	☐	☐	☐
9. ...hat erkannt, dass es sich bei Energietransferum die Weitergabe einer weniger anschaulichen Größe handelt, z. B. die Energie aus der Nahrung wird in chemischer Energie umgewandelt.	☐	☐	☐	☐
10. ...hat erkannt, dass die verschiedenen Energieformen nicht unabhängig voneinander sind, z. B. Lageenergie und Bewegungsenergie stehen immer in Verbindung und wandeln sich ineinander um (z. B. Pendel).	☐	☐	☐	☐
11. ...hat erkannt, dass Energieformen ineinander überführt werden können, z. B. Lageenergie wird in Bewegungsenergie umgewandelt oder thermische Energie wird in innere Energie umgewandelt.	☐	☐	☐	☐
12. ...hat erkannt, dass bei Energietransfer auch immer Wärmeenergie auftritt, z. B. bei der Umwandlung von chemischer Energie (Kohle) in Wärmekraftwerken, Abgabe der Energie als thermische Energie an die Umgebung..	☐	☐	☐	☐

→

	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
13. ...hat erkannt, dass es bei Energietransfer zu ungewollter Energiedissipation kommen kann (Wirkungsgrad), z. B. bei jedem Energietransport kann die Energie als thermische Energie an die Umgebung abgegeben werden.	☐	☐	☐	☐
14. ...hat erkannt, dass eine Energieform auf bestimmte Weisen genutzt werden können, z. B. Strahlungsenergie der Sonne kann beim Wachstum einer Weizenähre in chemische Energie umgewandelt werden, die als Bestandteil des Brotes dem Menschen chemische Energie liefert, die dann in Bewegungs- oder thermische Energie umgewandelt werden kann.	☐	☐	☐	☐
15. ...hat erkannt, dass Energieentwertung den Verlust nutzbarer Energie durch Umwandlung in innere Energie darstellt, z. B. elektrische Energie wird in einer Heizung als thermische Energie in der Umgebung abgegeben und ist somit nicht mehr nutzbar, entwertet..	☐	☐	☐	☐
16. ...hat erkannt, dass Energieentwertung nur in einer Richtung stattfindet (Irreversibilität), z. B. thermische Energie, die an die Umgebung abgegeben wird, kann nicht mehr genutzt werden (nicht mehr eingefangen werden und genutzt werden).	☐	☐	☐	☐
17. ...hat erkannt, dass die Gesamtenergie eines (geschlossenen) Systems erhalten bleibt und bilanziert wird, z. B. dass es Energieverbrauch im eigentlichen Sinne nicht gibt, da Energieerhaltung gilt und dass Energie nicht erzeugt und vernichtet werden kann.	☐	☐	☐	☐
18. ...kann die Notwendigkeit zum „Energiesparen“ begründen, z. B., dass die fossilen Brennstoffe, die zur Energiegewinnung benötigt werden, sich dem Ende neigen.	☐	☐	☐	☐

Studie 2 (gekürzt)

Bewerterkürzel:_____ MapID:_____ Rating_L_10 Items

Beurteilungsbogen

Liebe Lehrerin, lieber Lehrer.

Im Folgenden finden Sie Aussagen, die Sie bitte auf Grundlage der vorliegenden Concept Map bewerten. Bitte kreuzen Sie jede Aussage an, inwiefern sie für die Concept Map der Schülerin bzw. der Schülers zutrifft oder nicht.

Die Schülerin/ Der Schüler …

	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
1. …hat erkannt, dass Objekte bzw. Körper über Energie verfügen können, z. B. eine Batterie besitzt elektrische Energie.	☐	☐	☐	☐
2. …kann verschiedene Energiequellen benennen, z. B. Sonne als Energiequelle, Nahrung als Energiequelle.	☐	☐	☐	☐
3. …hat erkannt, dass die Energie der Objekte in Verbindung mit bestimmten Eigenschaften des Objektes bzw. Körpers steht, z. B. ruht ein Stein besitzt er Lageenergie, rollt er, besitzt er Bewegungsenergie.	☐	☐	☐	☐
4. …hat erkannt, dass Energie notwendig ist, um Arbeit zu verrichten, z. B. Energie ist die Fähigkeit, um Arbeit zu verrichten.	☐	☐	☐	☐
5. …hat erkannt, dass Energie in Energieträgern transportiert wird, z. B. Nahrung oder Kohle sind Energieträger.	☐	☐	☐	☐
6. …hat erkannt, dass die verschiedenen Energieformen nicht unabhängig voneinander sind, z. B. Lageenergie und Bewegungsenergie stehen immer in Verbindung und wandeln sich ineinander um (z. B. Pendel).	☐	☐	☐	☐
7. …hat erkannt, dass Energieformen ineinander überführt werden können, z. B. Lageenergie wird in Bewegungsenergie umgewandelt oder thermische Energie wird in innere Energie umgewandelt.	☐	☐	☐	☐
8. …hat erkannt, dass Energieentwertung den Verlust nutzbarer Energie durch Umwandlung in innere Energie darstellt, z. B. elektrische Energie wird in einer Heizung als thermische Energie in der Umgebung abgegeben und ist somit nicht mehr nutzbar, entwertet.	☐	☐	☐	☐
9. …hat erkannt, dass Energieentwertung nur in einer Richtung stattfindet (Irreversibilität), z. B. thermische Energie, die an die Umgebung abgegeben wird, kann nicht mehr genutzt werden (nicht mehr eingefangen werden und genutzt werden).	☐	☐	☐	☐
10. …hat erkannt, dass die Gesamtenergie eines (geschlossenen) Systems erhalten bleibt und bilanziert wird, z. B. dass es Energieverbrauch im eigentlichen Sinne nicht gibt, da Energieerhaltung gilt und dass Energie nicht erzeugt und vernichtet werden kann.	☐	☐	☐	☐
Jedes gesetzte Kreuz steht für eine Punktzahl (s.o.). Bitte addieren Sie die Punktzahlen.	Gesamtpunktzahl: _____ / 30			

A.3 Lehrerfragebogen zu Ausbildung und Beruf

Bewerterkürzel (Ihre Initialen):_____

Füllen *Sie aus bzw.* ***kreuzen*** *Sie die entsprechende Antwort an.*

1a. Seit wievielen Jahren unterrichten Sie? (Referendariatszeit ausgeschlossen, bitte in Jahren angeben)

_____**oder**

1b. Wenn Sie im Referendariat sind, in welchem Ausbildungshalbjahr befinden Sie sich?

_____

2. Welchen Studienabschluss haben Sie?

( ) Lehramt und 1. und 2. Staatsexamen und Referendariat
( ) Lehramt und 1. Staatsexamen und momentan noch im Referendariat
( ) Lehramt und 1. Staatsexamen und kein Referendariat
( ) Diplom und 2. Staatsexamen und Referendariat
( ) Diplom und kein 2. Staatsexamen und kein Referendariat
( ) Sonstiges:
_____

3. Wieviele Schülerinnen und Schüler der Lerngruppe hatten in den letzten Jahren bei Ihnen Physikunterricht und wie lange? Bitte tragen Sie die Anzahl der Schülerinnen und Schüler ein.

Schülerinnen und Schüler hatten bei mir vorher Physikunterricht:

_____ keinen.

_____ in einem Schulhalbjahr.

_____in zwei Schulhalbjahren.

_____in drei Schulhalbjahren.

_____in vier und mehr Schulhalbjahren.

4. Wie viele Schülerinnen und Schüler kennen Siebereits aus Unterricht aus Ihren anderen Fächern und wie lange? Bitte tragen Sie die Anzahl und das Fach ein.

Schülerinnen und Schüler hatten bei mir vorher in einem anderen Fach Unterricht:

_____ keinen.

_____ in einem Schulhalbjahrim _____unterricht.

_____ in zwei Schulhalbjahren im _____unterricht.

_____ in drei Schulhalbjahren im _____unterricht.

_____ in vier und mehr Schulhalbjahren im _____unterricht.

5. Welche Formen von Erhebungen kennen Sie (z. B. Münsteraner Rechtschreibanalyse, Tests…)?

6. Welche von diesen Formen nutzen Sie als diagnostische Instrumente in Ihrem Unterricht?

7. Kennen Sie Concept Mapping als Diagnoseinstrument?
( ) Ja
( ) Nein

8. Wie oft nutzen Sie Concept Mapping aktiv in Ihrem Unterricht? (Sie selber erstellen eine Concept Map bzw. lassen Ihre Schülerinnen und Schüler eine erstellen)

Nie – selten – manchmal – häufig – immer
( ) - ( ) - ( ) - ( ) - ( )

9. Kreuzen Sie bitte Ihr Geschlecht an.
( ) weiblich
( ) männlich

10. Geben Sie bitte Ihr Alter an.
_____ Jahre.

Vielen Dank für Ihre Teilnahme!

A.4 Manual zur Nutzung des Concept Map-Beurteilungsbogens

Gruppe 1

Manual zur Beantwortung der Items im Concept Map Beurteilungsbogen

Liebe Lehrerin, lieber Lehrer,

dies ist eine kurze Anleitung, wie Sie bei der Beurteilung der Concept Maps vorgehen sollen.

1. Jede Concept Map hat unten rechts bzw. oben links einen Code erhalten, z. B. 1-051-101.
2. Jede Concept Map wird mit einem Beurteilungsbogen beurteilt. Zunächst schreiben Sie oben links die Codenummer der Concept Map, die Sie gerade beurteilen. Im nächsten Schritt schreiben Sie bitte die Initialen Ihres Namen unter „Teilnehmernummer“, z. B. AM für Andreas Müller oben auf den Beurteilungsbogen.

Beispiel:

Bewerterkürzel: AM MapID: 1-051-101 Ratingsheet- Gesamtkonzept„Energie_L“

Beurteilungsbogen

Liebe Lehrerin, lieber Lehrer.

Im Folgenden finden Sie Aussagen, die Sie bitte auf Grundlage der Concept Map Ihrer Schülerin bzw. Ihres Schülers bewerten. Bitte kreuzen Sie für jede Aussage an, inwiefern sie für die Concept Map der Schülerin bzw. der Schülers zutrifft oder nicht.

Die Schülerin/ Der Schüler …

3. Nun können Sie mit der Bewertung beginnen.

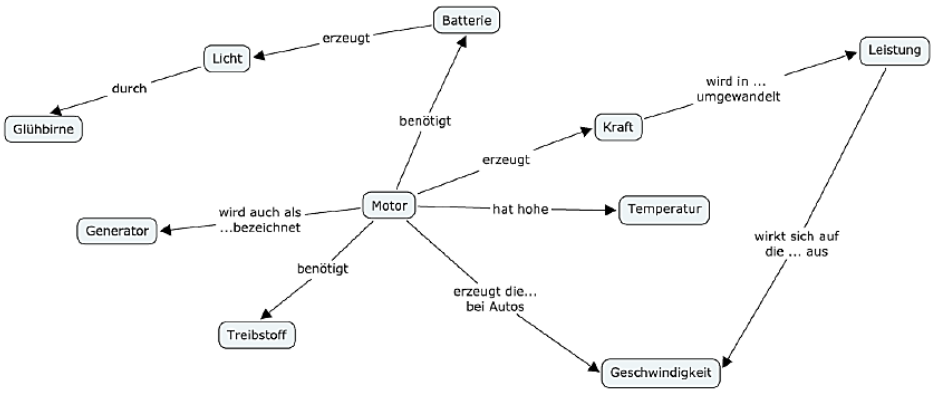
Wichtig für die Beurteilung:
Zwei verbundene Begriffe zusammen mit der bezeichneten Verbindung nennt man Proposition. Eine Concept Map besteht aus vielen Propositionen, die Sie aber nicht einzeln bewerten sollen. Vielmehr sollen Sie immer die gesamte Concept Map betrachten, um sich ein Urteil zu den abgefragten Aussagen zu machen. Es geht nicht darum, dass Sie gültige Propositionen zählen, sondern Sie sollen sich ein *ganzheitliches* Urteil bilden.
Dabei können Sie sich ungefähr an folgenden Stufen orientieren:

trifft nicht zu:	Das zu bewertende Kriterium ist nicht zu erkennen.
trifft nur ansatzweise zu:	Das zu bewertende Kriterium ist kaum zu erkennen, so dass eine Einschätzung des physikalischen Verständnisses nicht möglich ist.
trifft überwiegend zu:	Das zu bewertende Kriterium tritt so auf, dass ein physikalisch korrektes Verständnis vermutet werden kann.
trifft völlig zu:	Das zu bewertende Kriterium tritt so auf, dass ein physikalisch korrektes Verständnisses sichtbar wird.

Die nachfolgenden Beispiele sollen das anhand der ersten Frage verdeutlichen.

Kategorie: „trifft nicht zu“, es ist kein physikalisches Bild zu erkennen
0 Punkte:

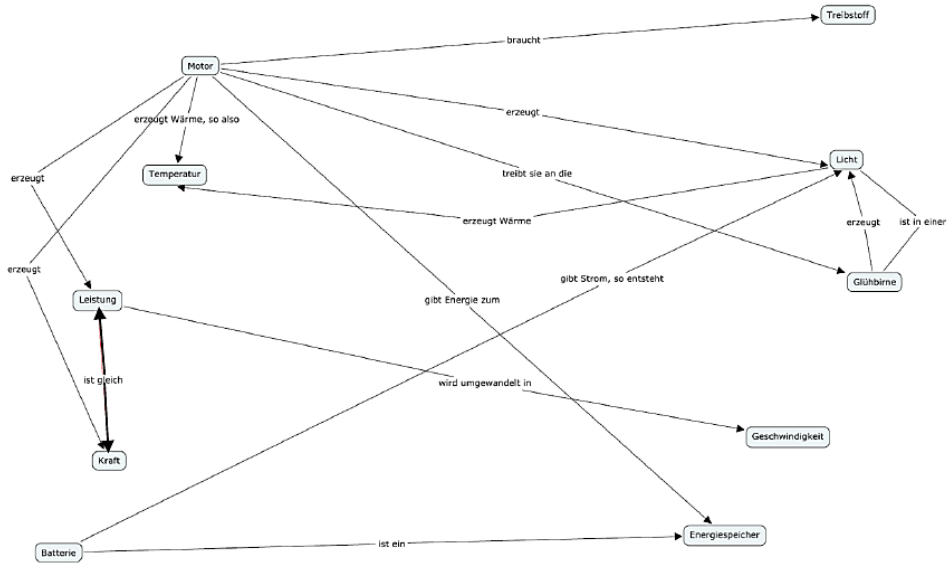
Die Schülerin/ Der Schüler…	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
1. …hat erkannt, dass Objekte bzw. Körper über Energie verfügen können, z. B. eine Batterie besitzt elektrische Energie.	☐	☐	☐	☐



Sie würden auf Basis dieser Map „trifft nicht zu“ ankreuzen, weil kein Bezug zur Energie eines Objekts hergestellt wird.

Kategorie: „trifft nur ansatzweise zu“, es ist ein falsches oder in Ansätzen richtiges physikalisches Bild zu erkennen
1 Punkt:

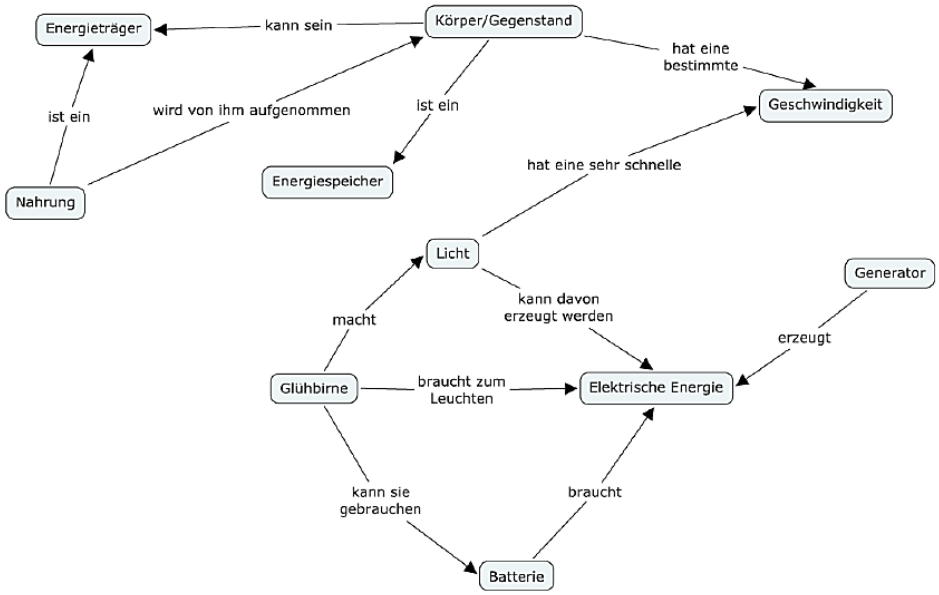
Die Schülerin/ Der Schüler…	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
1. …hat erkannt, dass Objekte bzw. Körper über Energie verfügen können, z. B. eine Batterie besitzt elektrische Energie.	☐	☐	☐	☐



Sie würden auf Basis dieser Map „trifft nur ansatzweise zu“ ankreuzen, weil nur vereinzelt ein Bezug zur Energie von Objekten („Motor gibt Energie zum Energiespeicher“ oder „Batterie ist ein Energiespeicher“.) hergestellt wird und eine Einschätzung über das physikalische Verständnis nicht möglich ist.

Kategorie: „trifft überwiegend zu“, es ist ein ungenaues, zum größten Teil aber richtiges physikalisches Bild zu erkennen
2 Punkte:

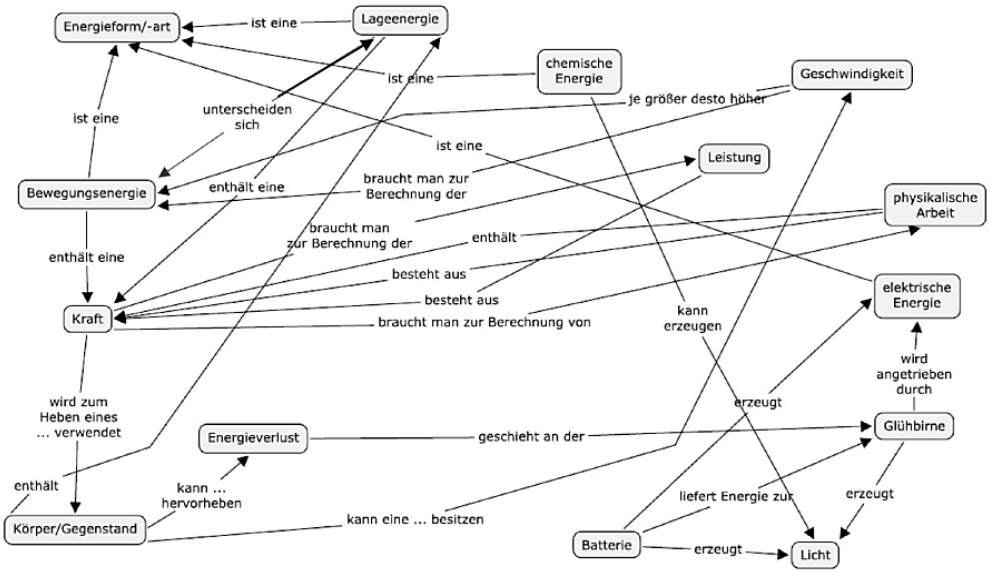
Die Schülerin/ Der Schüler…	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
1. …hat erkannt, dass Objekte bzw. Körper über Energie verfügen können, z. B. eine Batterie besitzt elektrische Energie.	☐	☐	☐	☐



Sie würden auf Basis dieser Map „trifft überwiegend zu“ ankreuzen, weil verschiedene Bezüge zur Energie von Objekten („Glühbirne braucht zum Leuchten Elektrische Energie“ oder „Batterie braucht Elektrische Energie“, „Nahrung ist Energieträger) hergestellt werden und ein physikalisch korrektes Verständnis vermutet werden kann.

Kategorie: „trifft völlig zu“, es ist ein genaues richtiges physikalisches Bild zu erkennen
3 Punkte:

Die Schülerin/ Der Schüler…	trifft völlig zu (3 Pkt.)	trifft überwiegend zu (2 Pkt.)	trifft nur ansatzweise zu (1 Pkt.)	trifft nicht zu (0 Pkt.)
1. …hat erkannt, dass Objekte bzw. Körper über Energie verfügen können, z. B. eine Batterie besitzt elektrische Energie.	☐	☐	☐	☐



Sie würden auf Basis dieser Map „trifft völlig zu“ ankreuzen, weil die Energie von Objekten (mehrfach) physikalisch korrekt („Körper enthält Lageenergie“ und „Glühbirne wird angetrieben durch elektrische Energie“) benannt wird.

Gruppe 3

Manual zur Beantwortung der Items im Beurteilungsbogen

Liebe Lehrerin, lieber Lehrer,

dies ist eine kurze Anleitung, wie Sie bei der Beurteilung Ihres Kursesvorgehen sollen.

1. Jede Schülerin/jeder Schüler hat von uns einen Code beim letzten Schulbesuch erhalten, z. B. 3-061-101.Nur Sie haben die Namen zu den Codes.Uns interessieren die Namen der Schülerinnen und Schüler nicht. Daher sollen Sie für den Datenschutz die Beurteilung über die Codes vornehmen.

2. Jede Schülerin/jeder Schüler wird mit einem Beurteilungsbogen beurteilt. Schreiben Sie oben auf jeden Beurteilungbogen bitte die Codenummer der Schülerin/des Schülers, die/den Sie gerade beurteilen und Ihr Bewerterkürzel. Das Bewerterkürzel ergibt sich aus den Initialen Ihres Namens, z. B. AM für Andreas Müller.

Beispiel:

Bewerterkürzel: AM SchülerID: 3-061-101 Ratings_L_10 Items

Beurteilungsbogen

Liebe Lehrerin, lieber Lehrer.

Bitte kreuzen Sie jede Aussage an, inwiefern sie für die zu beurteilende Schülerin bzw. den Schüler zutrifft oder nicht.

Die Schülerin/ Der Schüler ...

3. Bevor Sie mit der Beurteilung jeder einzelnen Schülerin/jedes einzelnen Schülers beginnen, beachten Sie bitte, dass Sie sich ein ganzheitliches Urteil über das momentane Verständnis Ihrer Schülerin/Ihres Schülers zum Basiskonzept Energie bilden sollen.
Hierzu lesen Sie zunächst die erste Aussage im Beurteilungsbogen durch und überlegen, inwieweit diese Aussage auf die betreffende Schülerin/den betreffenden Schüler zutrifft und kreuzen ein Feld an (entweder trifft nicht zu- trifft nur ansatzweise zu- trifft überwiegend zu oder trifft völlig zu). Was die Felder genau zu bedeuten haben, wird nachfolgendend beschrieben.

trifft nicht zu:	Das zu bewertende Kriterium ist nicht zu erkennen.
trifft nur ansatzweise zu:	Das zu bewertende Kriterium ist kaum zu erkennen, so dass eine Einschätzung des physikalischen Verständnisses nicht möglich ist.
trifft überwiegend zu:	Das zu bewertende Kriterium tritt so auf, dass ein physikalisch korrektes Verständnis vermutet werden kann.
trifft völlig zu:	Das zu bewertende Kriterium tritt so auf, dass ein physikalisch korrektes Verständnisses sichtbar wird.

Jede einzelne Aussage auf dem Bogen soll so für jede Schülerin/jeden Schüler beantwortet werden.

4. Nun können Sie mit der Bewertung beginnen.

A.5 Rankingbögen der verschiedenen Gruppen

Gruppe 1

CM-Ranking_01

Liebe Lehrerin, lieber Lehrer,

Sie haben bereits auf Basis der Concept Maps Ihrer Schülerinnen und Schüler den Beurteilungsbogen für jede Schülerin/jeden Schüler ausgefüllt. Nun sollen Sie die Beurteilungsbögen in eine Reihenfolge/Rangfolge bringen.

- Zunächst sortieren/gruppieren Sie die Beurteilungsbögen nach den Summenwerten. Dabei zeigen hohe Summenwerte ein breites Verständnis des Basiskonzepts Energie und niedrige Summenwerte wenig Verständnis des Basiskonzepts. Dies ist das erste Kriterium, an das Sie sich bei der Reihenfolgenbildung orientieren.
- Im nächsten Schritt verwenden Sie als weiteres Kriterium die inhaltliche Qualität der Beurteilungsbögen. Versuchen Sie dem Summenwert auch eine inhaltliche Dimension zu geben, indem Sie in den Beurteilungsbogen eines Schülers xy schauen. Z. B. können Sie dann sagen, dass der Schüler xy mit dem Summenwert 20 ein physikalisch fast korrektes Bild vom Basiskonzept Energie hat.
- Wenn Ihnen das klar ist, positionieren Sie denjenigen Beurteilungsbogen auf die Position 1, der Ihrer Meinung nach das meiste physikalische Verständnis hervorbringt.
- Wenn mehrere Schülerinnen und Schüler den gleichen Summenwert im Beurteilungsbogen erhalten haben, schauen Sie bitte erneut auf die inhaltliche Dimension in die betreffenden Beurteilungsbögen und versuchen Differenzen zwischen diesen Schülerinnen und Schülern festzustellen, um sie unter Umständen auf Basis der inhaltlichen Dimension auf verschiedene Positionen zu setzen.
- Natürlich ist es Ihnen erlaubt, auch mehrere Schülerinnen und Schüler auf eine Position zu setzen. Allerdings möchten wir Sie bitten, vorher noch einmal auf den Inhalt der Beurteilungsbögen zu schauen und Ihre Meinung evtl. doch zu revidieren.
- Halten Sie schließlich die von Ihnen erstellte Rangordnung auf dem angehefteten Zettel fest. Notieren Sie dazu jeweils den Rang und den Code der Schülerinnen und Schüler. Aus Datenschutzgründen sind wir nicht an den Namen der Schülerinnen und Schüler interessiert.

Vielen Dank!

Beispiel:

Rang	Code
1	1-051-105
2	1-051-110
2	1-051-108
3	1-051-130
...	...

CM-Ranking_01

Rang	Code
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	

Gruppe 2

CM-Ranking_02

Liebe Lehrerin, lieber Lehrer,

Sie sollen nun die Concept Maps Ihrer Schülerinnen und Schüler in eine Reihenfolge/Rangfolge bringen.

- Zunächst sortieren/gruppieren Sie die Concept Maps nach der inhaltlichen Qualität. Dabei können Sie beispielsweise feststellen, dass Schüler xy ein physikalisch fast korrektes Bild vom Basiskonzept Energie hat und deshalb gehört dieser Schüler Ihrer Meinung nach zu den Klassenstärkeren hinsichtlich des Verständnisses des Basiskonzepts Energie.
- Wenn Ihnen klar ist, welche Qualität die Concept Maps haben, positionieren Sie diejenige Concept Map auf die Position 1, die Ihrer Meinung nach das meiste physikalische Verständnis hervorbringt.
- Wenn mehrere Concept Maps die gleiche inhaltliche Qualität aufweisen, schauen Sie bitte erneut auf die inhaltliche Dimension der betreffenden Concept Maps und versuchen Differenzen zwischen diesen Concept Maps festzustellen, um sie unter Umständen auf Basis der inhaltlichen Dimension auf verschiedene Positionen zu setzen.
- Natürlich ist es Ihnen erlaubt, auch mehrere Concept Maps auf eine Position zu setzen. Allerdings möchten wir Sie bitten, vorher noch einmal auf den Inhalt der Concept Maps zu schauen und Ihre Meinung evtl. doch zu revidieren.
- Halten Sie schließlich die von Ihnen erstellte Rangordnung auf dem angehefteten Zettel fest. Notieren Sie dazu jeweils den Rang und den Code der Schülerinnen und Schüler (Concept Map). Aus Datenschutzgründen sind wir nicht an den Namen der Schülerinnen und Schüler interessiert.

Vielen Dank!

Wie in Gruppe 1 folgt die Tabelle, in der die Rangordnung eingetragen werden kann.

Gruppe 3

CM_BB-Ranking_03

Liebe Lehrerin, lieber Lehrer,

Sie haben bereits die Beurteilungsbögen für jede einzelne Schülerin/jeden einzelnen Schüler ausgefüllt. Nun sollen Sie die Beurteilungsbögen in eine Reihenfolge/Rangfolge bringen.

- Zunächst sortieren/gruppieren Sie die Beurteilungsbögen nach den Summenwerten. Dabei zeigen hohe Summenwerte ein breites Verständnis des Basiskonzepts Energie und niedrige Summenwerte wenig Verständnis des Basiskonzepts. Dies ist das erste Kriterium, an das Sie sich bei der Reihenfolgenbildung orientieren.
- Im nächsten Schritt verwenden Sie als weiteres Kriterium die inhaltliche Qualität der Beurteilungsbögen. Versuchen Sie dem Summenwert auch eine inhaltliche Dimension zu geben, indem Sie in den Beurteilungsbogen eines Schülers xy schauen. Z. B. können Sie dann sagen, dass der Schüler xy mit dem Summenwert 20 ein physikalisch fast korrektes Bild vom Basiskonzept Energie hat.
- Wenn Ihnen das klar ist, positionieren Sie denjenigen Beurteilungsbogen auf die Position 1, der Ihrer Meinung nach das meiste physikalische Verständnis hervorbringt.
- Wenn mehrere Schülerinnen und Schüler den gleichen Summenwert im Beurteilungsbogen erhalten haben, schauen Sie bitte erneut auf die inhaltliche Dimension in die betreffenden Beurteilungsbögen und versuchen Differenzen zwischen diesen Schülerinnen und Schülern festzustellen, um sie unter Umständen auf Basis der inhaltlichen Dimension auf verschiedene Positionen zu setzen.
- Natürlich ist es Ihnen erlaubt, auch mehrere Schülerinnen und Schüler auf eine Position zu setzen. Allerdings möchten wir Sie bitten, vorher noch einmal auf den Inhalt der Beurteilungsbögen zu schauen und Ihre Meinung evtl. doch zu revidieren.
- Halten Sie schließlich die von Ihnen erstellte Rangordnung auf dem angehefteten Zettel fest. Notieren Sie dazu jeweils den Rang und den Code der Schülerinnen und Schüler. Aus Datenschutzgründen sind wir nicht an den Namen der Schülerinnen und Schüler interessiert.- Sie haben bereits für jede Schülerin/jeden Schüler einen Beurteilungsbogen ausgefüllt.

Vielen Dank!

Wie in Gruppe 1 folgt die Tabelle, in der die Rangordnung eingetragen werden kann.

Gruppe 4

Ranking_04

Liebe Lehrerin, lieber Lehrer,

Sie sollen nun Ihre Schülerinnen und Schüler in eine Reihenfolge/Rangfolge bringen:

- Im Vorfeld hat jede Schülerin und jeder Schüler einen Code erhalten. Nur Sie haben die Codes mit den dazu gehörigen Namen.
- Bilden Sie sich bitte anhand der Namensliste eine Meinung über das Verständnis des Energiekonzepts der einzelnen Schülerinnen und Schüler.
- Ordnen Sie dann die Schülerinnen und Schüler nach Ihrer Einschätzung an. Beginnen Sie hierbei mit der Schülerin/dem Schüler, die/der nach Ihrer Meinung nach das Energiekonzept am besten verstanden hat. Diese Schülerin/dieser Schüler erhält die Position 1.
- Sie dürfen auch mehrere Schülerinnen und Schüler auf die gleiche Position setzen, wenn Sie der Meinung sind, dass die entsprechenden Schülerinnen und Schüler das gleiche Niveau haben.
- Halten Sie schließlich die von Ihnen erstellte Rangordnung auf dem angehefteten Zettel fest. Notieren Sie dazu jeweils den Rang und den Code der Schülerinnen und Schüler. Aus Datenschutzgründen sind wir nicht an den Namen der Schülerinnen und Schüler interessiert.

Vielen Dank!

Wie in Gruppe 1 folgt die Tabelle, in der die Rangordnung eingetragen werden kann.

B. Ergebnisse

Dieser Abschnitt zeigt die durchgeführten Analysen.

B1. Studie 1-nicht-parametrische Berechnungen

- *Konvergente Validität*

Tabelle B1.1. Korrelation nach Spearman zwischen Kompetenztest und Concept Map-Aufgabenformat allgemein, A und B.

Kompetenztest und Beurteilungsbogen	**Kompetenztest und Beurteilungsbogen bei Aufgabenformat A**	**Kompetenztest und Beurteilungsbogen bei Aufgabenformat B**
$\rho = .23^{*}, p < .05$	$\rho_A = .29, p = .07$	$\rho_B = .33^{*}, p < .05$

Bemerkungen: N_A = 40 Schülerinnen und Schüler, N_B = 39 Schülerinnen und Schüler. Signifikante Ergebnisse mit 5%iger Irrtumswahrscheinlichkeit werden mit * markiert ($p < .05$). 1%ige Irrtumswahrscheinlichkeit wird mit ** markiert ($p < .01$).

- *Gruppenunterschiede im Concept Mapping*

Tabelle B1.2. Gruppenvergleich im *U*-Test für unabhängige Stichproben.

		Aufgabenformat A^{+}	**Aufgabenformat B**$^{++}$
alle Rater	**Leistung im Kompetenztest**	$Md = 9.00$	$Md = 8.00$
	***U*-Test**	$U = 702, p = .44, z = -.770, \omega = .09, d = .24, 1\text{-}\beta = .18$	
	Leistung im CM-BB	$Md = 6.41$	$Md = 8.83$
	***U*-Test**	$U = 388, p < .01, z = -3.84, \omega = .43, d = .72, 1\text{-}\beta = .88$	

Bemerkungen: Die Analyse wird zwischen den Schülergruppen, die das Aufgabenformat A und Aufgabenformat B bearbeitet haben, durchgeführt. $^{+}N_A$ = 40 Schülerinnen und Schüler, $^{++}N_B$ = 39 Schülerinnen und Schüler. Cohens *d* wird mit den parametrisch ermittelten Mittelwerten der Gruppen in den entsprechenden Variablen ermittelt, um näherungsweise die Teststärke $1\text{-}\beta$ des U-Tests berechnen zu können. Die durch dieses Verfahren ermittelte Teststärke, sollte allerdings nur als Richtwert einer Mindestteststärke aufgefasst werden. Die Nutzung des parametrischen *d* führt zu einer Unterschätzung der wahren Teststärke (vgl. Rasch, Friese, Hofmann & Naumann, 2010).

B2. Studie 2-parametrische Berechnungen

- *Einfluss der Gruppenzugehörigkeit auf die Diagnosegenauigkeit*

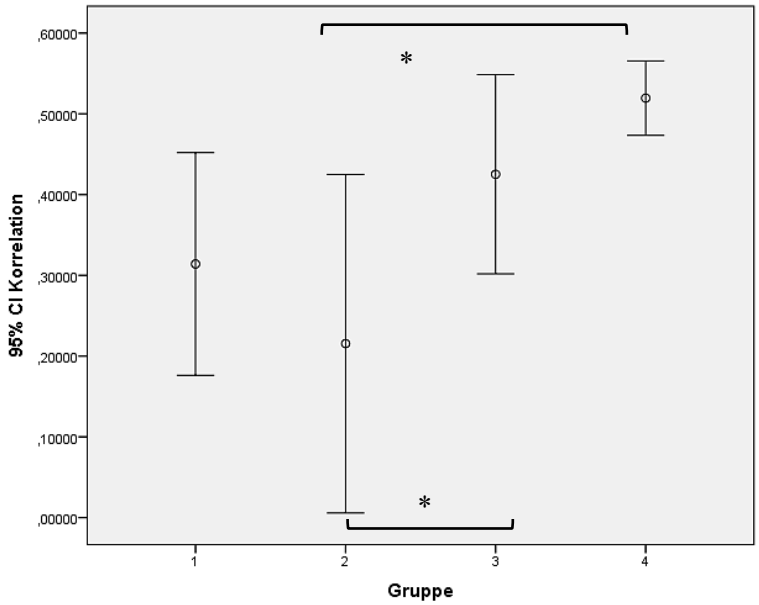


Abbildung B2.1. Mittelwerte der Rangkorrelationswerte (Diagnosegenauigkeit) bezogen auf die Gruppen (ANOVA).

Bemerkung: F (3,44) = 3.14, $p < .05$, $eta^2 = .18$. Es werden nur die signifikanten Ergebnisse markiert. * $p < .05$, ** $p < .01$. Fehler 95% Konfidenzintervall.

Tabelle B2.1. Mehrgruppenvergleiche im LSD-Post Hoc-Test bezogen auf die Diagnosegenauigkeit.

(I) Gruppe	(J) Gruppe	Mittlere Differenz (I-J)	Standardfehler	Signifikanz	Cohens *d*	1-β
1	2	.098	.096	.314	.30	.12
	3	-.111	.100	.276	.57	.28
	4	-.205	.109	.066	1.25	.78
2	3	-.210*	.100	.040	.73	.43
	4	-.304*	.107	.007	1.16	.74
3	4	-.094	.111	.400	.64	.28

Bemerkung: * $p < .05$, ** $p < .01$.

- *Einfluss der Kontrollvariablen auf die Schülerleistung im Kompetenztest und im Concept Mapping*

Tabelle B2.2. Korrelationsberechnungen nach Pearson für die Schülerstichprobe.

		Kompetenz-test	KFT	Physik-note	Mathe-note	Deutsch-note
Kompetenztest	*r*	1				
	Signifikanz					
	N	977				
KFT	*r*	.314**	1			
	Signifikanz	.000				
	N	971	971			
Physiknote	*r*	.381**	.266**	1		
	Signifikanz	.000	.000			
	N	964	958	964		
Mathenote	*r*	.343**	.242**	.599**	1	
	Signifikanz	.000	.000	.000		
	N	967	961	961	967	
Deutschnote	*r*	.103**	.120**	.378**	.407**	1
	Signifikanz	.001	.000	.000	.000	
	N	966	960	958	963	966

Bemerkung: * $p < .05$, ** $p < .01$.

Tabelle B2.3. Regressionsanalyse hinsichtlich der Kompetenztestleistung.

		B	***SE B***	β
Schritt 1	Konstante	4.789	.603	
	Physiknote	1.705	.134	.382**
	$R^2 = 0.146$			
Schritt 2	Konstante	1.754	.713	
	Physiknote	1.438	.135	.322**
	KFT	.232	.031	.226**
	$R^2 = 0.194$			
Schritt 3	Konstante	.854	.744	
	Physiknote	1.079	.163	.242**
	KFT	.218	.031	.213**
	Mathenote	.614	.157	.141**
	$R^2 = 0.206$			
Schritt 4 (Gesamtmodell)	Konstante	1.951	.823	
	Physiknote	1.165	.165	.261**
	KFT	.218	.031	.212**
	Mathenote	.730	.162	.168**
	Deutschnote	-.474	.164	-.093**
	$R^2 = 0.213$			

Bemerkungen: N_{gesamt} = 950 Schülerinnen und Schülern unter Berücksichtigung fehlender Daten. Der erklärende Anteil der Kontrollvariablen für die Kompetenztestleistung wird in R^2 ausgegeben. Für Schritt 1 beträgt der erklärende Anteil der Physiknote 0.146. 14.6% der Gesamtvariation der Kompetenztestleistung wird durch die Kontrollvariable Physiknote aufgeklärt. * $p < .05$, ** $p < .01$.

Die Schülerinnen und Schüler der Gruppe 1 erstellen Concept Maps. Inwiefern die KFT-Leistung mit der Leistung in den Concept Maps zusammenhängt, wird durch die Produkt-Moment-Korrelation nach Pearson exploriert. Die Punkte, die die Schülerinnen

und Schüler für ihre Concept Maps im Beurteilungsbogen erhalten haben, werden mit der Leistung im KFT in Bezug gesetzt. Der KFT korreliert mit der Concept Map-Bepunktung signifikant ($r = .23^{**}$, $p < .01$) und deutet darauf hin, dass die kognitiven Fähigkeiten einen geringen Zusammenhang mit der gezeigten Leistung im Concept Mapping haben.

- *Einfluss der Kontrollvariablen auf die Diagnosegenauigkeit*

Tabelle B2.4. Korrelationsberechnungen nach Pearson für die Lehrerstichprobe.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Diagnose-genauigkeit	*r*	1						
	Signifikanz							
	N	48						
(2) Gruppe	*r*	.322*	1					
	Signifikanz	.025						
	N	48	48					
(3) Anzahl der Berufsjahre	*r*	-.069	-.037	1				
	Signifikanz	.640	.802					
	N	48	48	48				
(4) Alter	*r*	.010	-.072	.913**	1			
	Signifikanz	.945	.632	.000				
	N	47	47	47	47			
(5) Geschlecht	*r*	.178	.361*	.109	.114	1		
	Signifikanz	.226	.012	.461	.446			
	N	48	48	48	47	48		
(6) Kenntnis von CM	*r*	.036	-.179	.225	.235	-.061	1	
	Signifikanz	.810	.229	.129	.116	.684		
	N	47	47	47	46	47	45	
(7) Nutzung von CM	*r*	.062	-.055	-.109	-.049	-.006	-.254	1
	Signifikanz	.688	.718	.476	.749	.968	.096	
	N	45	45	45	45	45	44	45

Bemerkungen: Die Variable Kenntnis von CM fragt die Lehrerinnen und Lehrer, ob sie Concept Mapping bereits kennen. Die Variable Nutzung von CM fragt die Lehrerinnen und Lehrer, inwiefern sie Concept Maps nutzen. Die genauen Fragen können im Anhang eingesehen werden.

Die Variablen Anzahl der Berufsjahre und das Alter korrelieren signifikant hoch. Diese Korrelation ist jedoch für die Beantwortung der Frage, welchen Einfluss die Kontrollvariablen auf die Diagnosegenauigkeit haben, irrelevant. Ebenso unbedeutend ist die signifikante Korrelation zwischen Gruppe und Geschlecht.

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

Tabelle B2.5. Kovarianzanalyse der Diagnosegenauigkeit.

Quelle	*F*	*df*	Signifikanz	*eta²*
Korrigiertes Modell	1.596	8	.162	.267
Konstanter Term	.744	1	.394	
Berufsjahre	2.034	1	.163	.043
Alter	1.389	1	.246	.030
Geschlecht	.447	1	.508	.009
Kenntnis von CM	1.265	1	.268	.026
Nutzung von CM	.156	1	.696	.003
Gruppe	3.293	3	.032	.210
Fehler		35		
R^2 = .267 (korrigiertes R^2 = .100), df_{gesamt} = 44				

Bemerkung: Vollständiger Datensatz von N = 44 Lehrerinnen und Lehrern.

- *Einfluss der Kontrollvariablen auf die Diagnosegenauigkeit, Einzelberechnungen*

Tabelle B2.6. Kovarianzanalyse der Diagnosegenauigkeit, Variable Berufsjahre.

Quelle	*F*	*df*	Signifikanz	*eta²*
Korrigiertes Modell	2.412	4	.064	.183
Konstanter Term	54.125	1	.000	
Berufsjahre	.352	1	.556	.006
Gruppe	3.131	3	.035	.179
Fehler		43		
R^2 = .183 (korrigiertes R^2 = .107), df_{gesamt} = 48				

Tabelle B2.7. Kovarianzanalyse der Diagnosegenauigkeit, Variable Alter.

Quelle	*F*	*df*	Signifikanz	*eta²*
Korrigiertes Modell	2.493	4	.057	.193
Konstanter Term	4.795	1	.034	
Alter	.069	1	.794	.001
Gruppe	3.323	3	.029	.192
Fehler		42		
R^2 = .192 (korrigiertes R^2 = .115), df_{gesamt} = 47				

Tabelle B2.8. Kovarianzanalyse der Diagnosegenauigkeit, Variable Geschlecht.

Quelle	*F*	*df*	Signifikanz	*eta²*
Korrigiertes Modell	2.418	4	.063	.184
Konstanter Term	1.014	1	.320	
Geschlecht	.375	1	.544	.007
Gruppe	2.667	3	.060	.152
Fehler		43		
R^2 = .184 (korrigiertes R^2 = .108), df_{gesamt} = 48				

Tabelle B2.9. Kovarianzanalyse der Diagnosegenauigkeit, Variable Kenntnis von Concept Maps.

Quelle	*F*	*df*	Signifikanz	*eta*^{2}
Korrigiertes Modell	2.457	4	.060	.190
Konstanter Term	2.077	1	.157	
Kenntnis von CM	.755	1	.390	.015
Gruppe	3.253	3	.031	.188
Fehler		42		
R^2 = .190 (korrigiertes R^2 = .112), df_{gesamt} = 47				

Tabelle B2.10. Kovarianzanalyse der Diagnosegenauigkeit, Variable Nutzung von Concept Maps.

Quelle	*F*	*df*	Signifikanz	*eta*^{2}
Korrigiertes Modell	2.301	4	.120	.187
Konstanter Term	14.547	1	.550	
Nutzung von CM	.194	1	.682	.004
Gruppe	3.005	3	.020	.183
Fehler		40		
R^2 = .187 (korrigiertes R^2 = .106), df_{gesamt} = 45				

- *Haupteffekte, Interaktion und Interaktionseffekt der eingesetzten Instrumente*

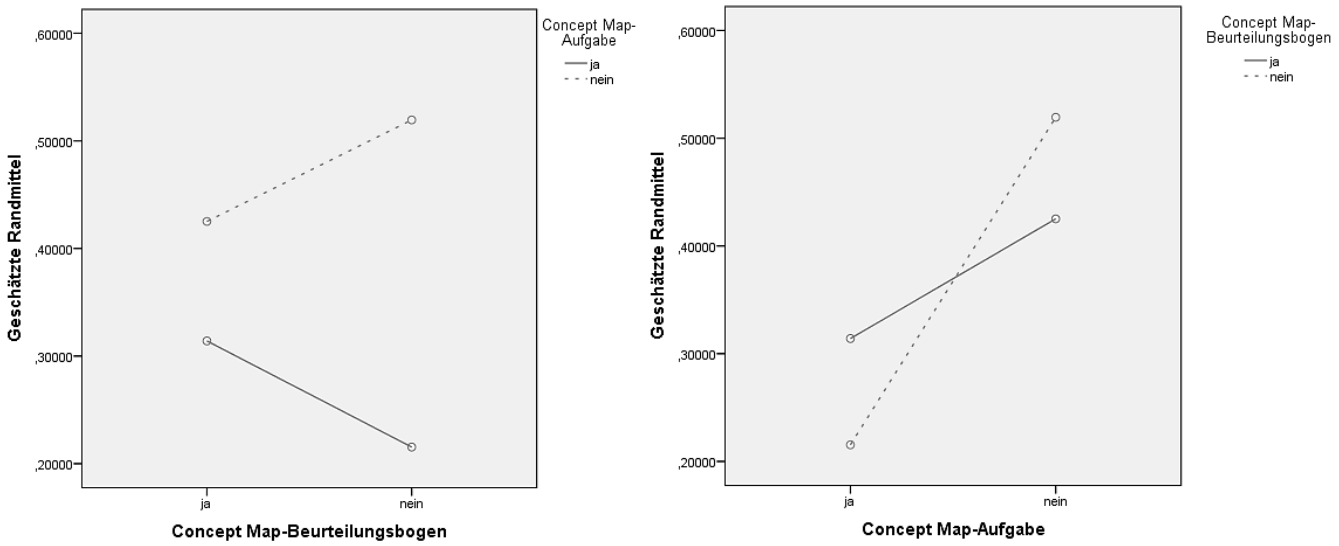


Abbildung B2.2. Hybride Interaktion zwischen Concept Map-Aufgabenformat und (Concept Map-) Beurteilungsbogen auf die mittleren Korrelationswerte (Diagnosegenauigkeit).

links: Profilplot Haupteffekt Concept Map-Beurteilungsbogen. rechts: Profilplot Haupteffekt Concept Map-Aufgabenformat.

Bemerkungen: Es wird dann von einer Interaktion gesprochen, wenn die Linien nicht parallel verlaufen (vgl. Sedlmeier & Renkewitz, 2008). Nach Bortz und Schuster (2010) kann diese Interaktion als hybride Interaktion klassifiziert werden.

Tabelle B2.11. Zweifaktorielle Varianzanalyse der Diagnosegenauigkeit.

Quelle	***F***	***df***	**Signifikanz**	***eta*²**
Korrigiertes Modell	3.144	3	.034	.177
Konstanter Term	100.262	1	.000	
Concept Map-Aufgabenformat	7.950	1	.007	.149
Concept Map-Beurteilungsbogen	.001	1	.976	.000
CM-Aufgabenformat * CM-Beurteilungsbogen	1.717	1	.197	.032
Fehler		44		
R^2 = .177 (korrigiertes R^2 = .120), df_{gesamt} = 48				

Bemerkungen: Die Tabelle zeigt einen signifikanten Haupteffekt des Concept Map-Aufgabenformats auf die Diagnosegenauigkeit, F (1,44) = 7.95, $p < .01$, eta^2 = .15. Der Beurteilungsbogen hat keinen signifikanten Einfluss, F (1,44) = .001, p = .98, eta^2 = .00. Die Kombination aus beiden Instrumenten hat ebenfalls keinen signifikanten Einfluss, F (1,44) = 1.72, p = .20, eta^2 = .03. An dieser Stelle darf nicht der Trugschluss entstehen, dass eine vorhandene Interaktion im Profilplot signifikant sein muss. Ob die Interaktion im Test signifikant wird, hängt von dem Grad ab, wie stark die Linien nicht parallel verlaufen (vgl. Field, 2009).

Publikationsliste

Die folgende Publikationsliste enthält sämtliche Veröffentlichungen, die im Rahmen dieser Arbeit und anderer Projekte entstanden sind.

2010

- Ley, S. L.: *Ein Vergleich von Schülervorstellungen in Nature of Science und Scientific Inquiry.* Unveröffentlichte Staatsexamensarbeit. Universität Duisburg-Essen: Essen.

2012

- Ley, S. L., Krabbe, H. & Fischer, H. E. Konvergente Validität von Concept Maps: Einsatz verschiedener Concept Mapping Aufgabenformate zur Schülerdiagnose im Physikunterricht im Vergleich zu einem Kompetenztest. In: S. Bernholt (Hrsg.): Konzepte fachdidaktischer Strukturierung für den Unterricht. *Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Oldenburg 2011* (S. 376-378). Berlin*: LIT-Verlag*.

- Ley, S. L., Krabbe, H. & Fischer, H. E. Convergent Validity: Concept Maps and Competence Test for students' Diagnosis in Physics. In: A. J. Cañas, J. D. Novak & J. Vanhear (Hrsg.): *Concept Maps: Theory, Methodology, Technology. Proc. of the Fifth Int. Conference on Concept Mapping* (S. 149-155). Malta: Veritas Press.

2014

- Ley, S. L., Krabbe, H. & Fischer, H. E. (in Vorb.). Schülerdiagnose durch Concept Maps. Ein Weg Schülerinnen und Schüler zu diagnostizieren. In: *Praxis in den Naturwissenschaften*. *Physik in der Schule*.

- Won, M., Ley, S. L., Krabbe, H., Treagust, D. & Fischer, H. E. (in Vorb.). Concept Maps as a formative assessment tool for the concept of energy.

Beiträge zu Konferenzen und Workshops

2011

- Poster auf dem nwu-Workshop, Essen, Deutschland:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Concept Maps als Diagnoseinstrument im Physikunterricht und die Messung der Diagnosegenauigkeit von Physiklehrkräften*.

- Poster auf der Summerschool des Joint Researcher Trainingsmit dem finnischen und niederländischen Graduiertenkolleg, Joensuu, Finnland:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Concept Maps as diagnostic instrument in Physics and the measurement of physics teachers diagnostic accuracy*.

- Vortrag auf der GDCP-Jahrestagung, Oldenburg, Deutschland:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Konvergente Validität von Concept Maps: Einsatz verschiedener Concept Mapping Aufgabenformate zur Schülerdiagnose im Physikunterricht im Vergleich zu einem Kompetenztest*.

- Vortrag auf der Winterschool des Joint Researcher Training mit dem finnischen und niederländischen Graduiertenkolleg, Hamburg, Deutschland:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Convergent Validity: Concept Maps and Competence Test for Students' Diagnosis in Physics*.

2012

- Workshopgestaltung auf dem Physiklehrertag NRW, Kamen, Deutschland:
  Krabbe, H. & Ley, S. L. *Diagnose mit Concept Maps*.

- Vortrag auf dem Kongress der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), Osnabrück, Deutschland:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Validierung eines Diagnoseinstrumentes für den Physikunterricht*.

- Vortrag auf der 5th international Conference on Concept Mapping, Valletta, Malta:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Convergent Validity: Concept Maps and Competence Test for Students' Diagnosis in Physics*.

- Poster auf der GDCP-Jahrestagung, Hannover, Deutschland:
  Krabbe, H., Ley, S. L. & Fischer, H. E. *Lernstandsdiagnostik mit Modalnetzen*.

2013

- Poster auf der nwu-Abschlussveranstaltung, Essen, Deutschland:
  Ley, S. L. *Concept Maps als Diagnoseinstrument im Physikunterricht und die Messung der Diagnosegenauigkeit von Physiklehrkräften*.

- Vortrag im Doctoral Colloquium des Science and Mathematics Education Centre (SMEC) der Curtin University of Technology, Perth, Australien:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Concept Maps as Diagnostic Instrument and their effect on Teachers' Diagnostic Accuracy in Physics*.

- Vortrag auf der ASERA 2013 Conference, Wellington, Neuseeland:
  Won, M., Ley, S. L. & Treagust, D. F. *Concept Maps as a diagnostic tool for teaching and learning physics*.

- Poster auf dem Internationalen Sommerfest der Universität Duisburg-Essen, Essen, Deutschland:
  Ley, S. L., Krabbe, H., Fischer, H. E., Won, M. & Treagust, D. F. *Concept Maps as a Diagnostic Tool for Teaching and Learning Physics*.

- Vortrag auf der ESERA 2013 Conference, Nicosia, Zypern:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Physics Teachers' Diagnostic Accuracy Using Concept Maps*.

- Vortrag auf der GDCP-Jahrestagung, München, Deutschland:
  Ley, S. L., Krabbe, H. & Fischer, H. E. *Diagnosegenauigkeit von Physiklehrkräften im Einsatz von Concept Maps*.

2014

- Vortrag auf der NARST 2014 Conference, Pittsburgh, USA:
  Krabbe, H., Ley, S. L. & Fischer, H. E. *Physics Teachers' Diagnostic Accuracy in the Use with Concept Maps*.

Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Danksagung

Ich möchte mich herzlich bei allen, die mich bei meiner Arbeit unterstützt haben, bedanken.

Mein erster Dank gilt **Prof. Dr. Hans E. Fischer**, meinem Doktorvater. Durch diese Arbeit ist mir ein Weg eröffnet worden, der mich um viele Erfahrungen reicher gemacht hat. Ich danke Hans für sein immer offenes Ohr, für diese wertvolle Zeit und seiner Persönlichkeit, die ich sehr zu schätzen weiß.

Prof. Dr. Helmut Fischler danke ich für die Begutachtung dieser Arbeit. Ihm und **Prof. Dr. Elke Sumfleth** danke ich zudem für die mündliche Prüfung, die mir positiv in Erinnerung bleiben wird. Elke danke ich auch für die gute Zusammenarbeit während meiner Zeit in der Forschergruppe.

Prof. Dr. Andreas Wucher danke ich für die Übernahme des Prüfungsvorsitzes und ebenfalls der angenehmen mündlichen Prüfung.

Dr. Heiko Krabbe danke ich herzlich für die sehr gute Mitbetreuung meiner Promotionszeit. Er hat großen Anteil daran, dass wir mein Projekt zu einem Projekt gemacht haben. Danke für die vielen fruchtvollen Diskussionen, die mir lange positiv in Erinnerung bleiben werden.

Prof. Dr. David Treagust, **Dr. Mihye Won** and the whole **SMEC-Team** at Curtin University, I would like to thank you for the wonderful time in Perth and the pleasant time to work with you during our German-Australian cooperation.

Dr. Irene und **Prof. Dr. Knut Neumann** danke ich für den „ersten Kontakt" mit der Wissenschaft und die bereichernde Arbeitszeit.

Dr. Tobias Viering danke ich für die Bereitstellung seines Kompetenztestes, ohne den dieses Projekt nur halb durchführbar gewesen wäre.

Meiner Arbeitsgruppe und den **Kolleginnen und Kollegen aus der Forschergruppe und dem Graduiertenkolleg** danke ich für die fachliche Unterstützung und den interessanten Diskussionen.

Prof. Dr. Detlev Leutner, Dr. Maria Opfermann, Dr. Annett Schmeck, Christian Spoden und **Benjamin Klein** danke ich für die „psychologisch"-statistische Beratung. Sie haben mir in statistischen Fragen, wo meine Expertise am Ende war, helfen können.

Ich danke den **studentischen Mitarbeitern**, die mich in vielen Dingen stark entlastet haben: **Tolga Artkan, Laura Ostermann, Hamid Rafiq, Roman Lettmann, Aynur Yüksel, Dominik Bures** und **Jens Kraft**. Sie haben sich für mein Projekt eingesetzt und einen

großen Beitrag zum Gelingen geleistet. **Jonathan Higgins** danke ich für die stete Englisch-Hilfe. **Claudia Evers, Janina Kubon, stellvertretend für das Videolabor Heiner Herriger** und **Hermann Vielhauer** danke ich für die immer reibungslos funktionierende Infrastruktur. Das Gleiche gilt für **Dr. Helene Kruse, Raffaela Römer** und **Sara Münzberg**. Ohne sie hätte ich nie so viele Projekteilnehmer gefunden. Lieben Dank!

Allen teilnehmenden **Lehrerinnen und Lehrern** und **Schülerinnen und Schülern,** im Besonderen **Udo Wlotzka,** bin ich zutiefst dankbar, dass sie mein Projekt angenommen haben wie es ist und mir somit die Möglichkeit gegeben haben, es durchzuführen. Für Ihre Hilfe und die Kooperationsbereitschaft danke ich.

Ich habe durch meine Arbeit viele schöne Gespräche, nicht nur fachlicher Natur führen können, und durfte Menschen kennenlernen, die ich sehr zu schätzen gelernt habe. **Prof. Dr. Markus Emden** und **Prof. Dr. Hendrik Härtig**. Lieben Dank fürs Zuhören.

Meike Bergs, Andreas Dickhäuser, Cornelia Geller, Nicola Großebrahm, Dominique Klein, Mirko Krüger, Manuela Lehnen, Stefan Mutke, Tobias Pollender, Norman Riehs, Nils Rohde, Maike Schmidt und **Felix Schoppmeier** sind ebenso Teil meiner Erinnerungen, die ich nicht missen möchte. Ich danke ihnen herzlichst, dass sie mich aufgenommen haben und für die tolle Atmosphäre. Ich werde mich gerne an ihre Persönlichkeiten zurückerinnern.

Bettina Kreiter möchte ich darüber hinaus für die schöne Bürogemeinschaft und alles drum herum während meiner Arbeit danken. Es war mir immer eine Freude mit ihr!

Meinen Freunden und **Josef Riese** möchte ich für die Zeiten abseits der Arbeit danken.

Ein Büro ist kein Büro, wenn es keine Menschen beinhalten würde. Mit meinem Büropartner und gutem Freund **Simon Zander** habe ich so manch schöne Zeit in diesem Büro verbracht. Ich danke ihm, dass er immer für mich da war und ich werde unsere Gespräche rund um das Leben und die Arbeit in guter Erinnerung haben.

Familie Krumme danke ich ebenfalls für die immer herzliche Unterstützung!

Den Schluss dieser Danksagung widme ich **meiner Familie**. Meinen **Eltern** und meinen **Geschwistern und ihren Familien** und meinem **Onkel, seiner Frau und ihren Kindern** möchte ich für die Zeiten außerhalb der Arbeit und ihren ganz eigenen Unterstützungsstrategien danken. Ohne sie geht es nicht.

Mein letzter Dank gilt **Bernhard**. Sein Verständnis, sein Zuspruch und sein Lachen haben mich weitermachen lassen und unsere Beziehung weiter gestärkt. Er hat immer an mich geglaubt und mich unterstützt.

Ich danke allen für die Unterstützung!

Erklärung

Ich versichere, dass ich die eingereichte Dissertation selbstständig verfasst habe.

Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Alle Stellen und Formulierungen, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht.

Essen, den 08. Oktober 2014 (Siv Ling Ley)