Bewertungskompetenz im Physikunterricht: Entwicklung eines Messinstruments zum Themenfeld Energiegewinnung, -speicherung und -nutzung

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen

im Promotionsprogramm *ProPhys*
der *Georg-August University School of Science (GAUSS)*

vorgelegt von

Mark Torsten Sakschewski
aus Walsrode

Göttingen, 2013

Betreuungsausschuss:

Prof. Dr. Susanne Schneider
Didaktik der Physik, IV. Physikalisches Institut, Fakultät für Physik

Prof. Dr. Susanne Bögeholz
Didaktik der Biologie, Albrecht-von-Haller-Institut für Pflanzenwissenschaften,
Fakultät für Biologie und Psychologie

Prof. Dr. Stefan Dreizler
Sonnenphysik und Stellare Astrophysik, Institut für Astrophysik, Fakultät für Physik

Mitglieder der Prüfungskommission:

Referentin: Prof. Dr. Susanne Schneider
Didaktik der Physik, IV. Physikalisches Institut, Fakultät für Physik

Korreferentin: Prof. Dr. Susanne Bögeholz
Didaktik der Biologie, Albrecht-von-Haller-Institut für Pflanzenwissenschaften,
Fakultät für Biologie und Psychologie

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Stefan Dreizler
Sonnenphysik und Stellare Astrophysik, Institut für Astrophysik, Fakultät für Physik

Prof. Dr. Stefan Halverscheid
Didaktik der Mathematik, Mathematisches Institut, Fakultät für Mathematik und Informatik

Prof. Dr. Wolfram Kollatschny
Extragalaktische Astrophysik und Kosmologie, Institut für Astrophysik, Fakultät für Physik

Jun.-Prof. Dr. Thomas Waitz
Abteilung für Fachdidaktik Chemie, Institut für Anorganische Chemie, Fakultät für Chemie

Tag der mündlichen Prüfung: 30. Oktober 2013

Inhaltsverzeichnis

Zusammenfassung

Die vorliegende Studie diskutiert die Entwicklung eines Testinstruments zur Messung von *Bewertungskompetenz* im Sinne der Teilkompetenz *Bewerten, Entscheiden und Reflektieren (BER)* innerhalb des *Göttinger Modells der Bewertungskompetenz im Kontext nachhaltiger Entwicklung* (Bögeholz 2011) für das Unterrichtsfach Physik in der Sekundarstufe.

Die ausgewählten Aufgabenkontexte beschreiben die Erzeugung, die Speicherung und die Nutzung elektrischer Energie. Sie schließen damit auch an die aktuelle gesellschaftliche Diskussion um Erneuerbare Energien an und untersuchen diesbezüglich das Entscheidungsvermögen und die Bewertungskompetenz heutiger Schülerinnen und Schüler.

Die Einsatzfähigkeit des in dieser Studie entwickelten Testinstruments wurde zunächst im Rahmen zweier Vorstudien überprüft, bevor die Haupterhebung als Querschnittstudie in den Jahrgängen 6, 8, 10 und 12 erfolgte ($N = 850$ Schülerinnen und Schüler an Gymnasien). Nach dem Ansatz von Eggert (2008), Eggert und Bögeholz (2006, 2010) ist es dabei als *paper-and-pencil*-Test konzipiert und beinhaltet zwei Entscheidungsaufgaben und eine Reflexionsaufgabe. Die empirisch gewonnenen Daten wurden zunächst anhand eines entwickelten Scoring Guides codiert und anschließend sowohl unter Gesichtspunkten der Klassischen als auch der Probabilistischen Testtheorie ausgewertet. Das entwickelte Testinstrument hat sich unter Reliabilitäts- und Validitätsaspekten bewährt. Item-Fit-Parameter zeigen, dass sich die empirischen Daten gut in einem eindimensionalen Rasch-Partial-Credit-Modell abbilden lassen.

Unter anderem konnten Zusammenhänge von BER mit dem Schulalter der Schülerinnen und Schüler nachgewiesen werden. Geringe Korrelationen von BER bestehen zu verschiedenen Schulnoten (u. a. zu Deutsch, Mathematik, Politik und Physik in der Klasse 10), zudem wird das Testergebnis für BER kaum von Lesekompetenzen beeinflusst.

Sakschewski, Mark (2013). Bewertungskompetenz im Physikunterricht: Entwicklung eines Messinstruments zum Themenfeld Energiegewinnung, -speicherung und -nutzung. Dissertation, Fakultät für Physik, Georg-August-Universität Göttingen. Verfügbar unter: http://hdl.handle.net/11858/00-1735-0000-0023-9918-8

Abstract (english)

The aim of this study was the development of a questionnaire, able to assess students' *decision-making competencies* related to physics education at upper secondary schools.

The chosen contexts deal with the generation, storage and use of electric energy and thereby tie up to public discussions about the chances and challenges of renewable energy sources.

The usability of the questionnaire was determined in two pilot studies, before the main study was conducted ($N = 850$, cross-sectional study, grades 6, 8, 10 and 12 at German "Gymnasium", the most academic track). Based on the approach of Eggert (2008), Eggert und Bögeholz (2006, 2010), a *paper-and-pencil test* was developed that contains two decision-making tasks and a reflection task. In order to analyse students' answers, a scoring guide was developed and used to evaluate data with regard to classical and probabilistic test theory. The developed measurement instrument is discussed with respect to reliability und validity aspects. Item fit statistics reveal that the final questionnaire fits the requirements of the Rasch partial credit model.

Relations between students' ability measures and years of education were detected as well as weak correlations between students' ability measures and some subject grades (German, Mathematics, Politics and Physics in grade 10), furthermore students' ability is only marginally affected by reading competencies.

Sakschewski, Mark (2013). Bewertungskompetenz im Physikunterricht: Entwicklung eines Messinstruments zum Themenfeld Energiegewinnung, -speicherung und -nutzung. Dissertation, Fakultät für Physik, Georg-August-Universität Göttingen.
Available online (german): http://hdl.handle.net/11858/00-1735-0000-0023-9918-8

Einleitung

Ziel

Diese Arbeit hat es sich zum Ziel gesetzt, das von der Göttinger Biologiedidaktik-Arbeitsgruppe entwickelte *Göttinger Modell der Bewertungskompetenz im Kontext nachhaltiger Entwicklung* (Bögeholz 2011, Eggert 2008, Eggert & Bögeholz 2006) auf das Unterrichtsfach Physik zu übertragen und aufzuzeigen, dass es auch im Rahmen des Physikunterrichts möglich ist, Bewertungskompetenz von Schülerinnen und Schülern nach dem Göttinger Modell zu modellieren.

Nationale Bildungsstandards

Begründet durch das nur mittelmäßige Abschneiden deutscher Schülerinnen und Schüler bei internationalen Schulleistungsstudien, ist das deutsche Bildungssystem stärker in das Bewusstsein der Menschen geraten. Es erfährt seit dem „PISA-Schock“ auch eine starke mediale Präsenz, die es Teil großer öffentlicher Debatten werden ließ und auch Bildungspolitikerinnen und Bildungspolitiker zum schnellen Handeln veranlasste. Die im Jahr 2004 von der Kultusministerkonferenz (KMK) verabschiedeten Nationalen Bildungsstandards sind eine Reaktion auf diesen Umstand. Neben Deutsch, Mathematik und erster Fremdsprache bedeutet die Einführung Nationaler Bildungsstandards auch einen stärkeren Fokus auf die Naturwissenschaften im Schulunterricht (KMK 2005a, b, c). Dabei wird angestrebt, den bis dahin oftmals vorherrschenden traditionellen Unterricht, der primär auf die reine Vermittlung von Fachwissen ausgelegt war, zunehmend durch einen solchen zu ersetzen, der durch gezielte Bezüge zur Gesellschaft und zum Alltag der Schülerinnen und Schüler auch den lebenspraktischen Nutzen von Naturwissenschaften betont (KMK 2005d). Bewertungskompetenzen sind in den Nationalen Bildungsstandards für das Unterrichtsfach Physik als eine von vier Kompetenzbereichen beschrieben (KMK 2005c).

fehlende Testinstrumente

Neben den Kompetenzbereichen *Fachwissen*, *Erkenntnisgewinnung* und *Kommunikation* kommt so auch dem Kompetenzbereich *Bewertung* eine wichtige Aufgabe zu (KMK 2005c). Derzeit sind zwar Standards für den Kompetenzbereich *Bewertung* beschrieben, deren Erreichung durch Schülerinnen und Schüler ließ sich bisher mangels geeigneter Diagnoseinstrumente aber kaum überprüfen. Fachdidaktische Forschungen zum Erreichen dieser Bildungsstandards und zu deren Evaluation konnten jedoch erst im Nachhinein

erfolgen. Hierzu bedarf es unter anderem qualifizierter Testinstrumente, die die gewünschten Kompetenzen bei Schülerinnen und Schülern erfassen, um diese dann auch modellieren zu können. Derzeit liegen nur vereinzelt Studien vor, die sich gezielt der Modellierung von Bewertungskompetenz in den drei Naturwissenschaften widmen (u. a. Eggert 2008, Eggert & Bögeholz 2006, Heitmann 2012, Hostenbach 2011, Knittel & Mikelskis-Seifert 2013).

Energie im Alltag

Vor dem Hintergrund des Leitbildes einer nachhaltigen Entwicklung (WCED 1987) ist das entwickelte Testinstrument konzipiert für physikalische Kontexte, in denen Erzeugung elektrischer Energie aus Windenergie, Speicherung elektrischer Energie und Nutzung von elektrischer Energie eine Rolle spielen. Energie hat einen bedeutenden Stellenwert in unserem Alltag und ist daher als Aufgabenkontext im besonderen Maße geeignet. Neben thermischer Energie für Heizung und Verkehr ist vor allem elektrische Energie für Menschen auf der ganzen Welt von essentieller Bedeutung. Das Thema Energie durchzieht also alle Facetten unserer Gesellschaft und hat damit eine direkte lebenspraktische Relevanz (Chedid 2005).

Bildungsauftrag

Im Sinne der Ausbildung von mündigen, reflektiert entscheidenden Bürgerinnen und Bürgern ist es daher Aufgabe von Bildungspolitik und Gesellschaft, Schülerinnen und Schüler in ihrem eigenen Interesse und im Interesse ihrer Nachfahren in diese Problematik einzuführen und sie für nachhaltiges Handeln zu sensibilisieren (Bögeholz 2006, Eggert & Hößle 2006, KMK 2005d). Dies schließt informiertes und begründetes Entscheiden mit ein. Ein reflektierter Umgang mit Entscheidungssituationen ist ein elementarer Bestandteil von Bewertungskompetenz und spielt daher auch in dem hier vorgestellten Testinstrument eine zentrale Rolle.

KAPITEL 1

Grundlagen und Einbettung des Forschungsvorhabens

Dieses Kapitel soll zunächst einige generelle Grundlagen und Begriffe einführen, auf denen aufbauend dann in Kapitel 2 der Stand der Forschung zu Kompetenzen, Bewertungskompetenz und das *Göttinger Modell der Bewertungskompetenz im Kontext nachhaltiger Entwicklung* erläutert wird.

1.1 Entscheidungen

permanent Entscheidungssituationen

Wir sind permanent Entscheidungssituationen ausgesetzt: „Schreibe ich jetzt noch weiter an meiner Dissertation oder gehe ich heute früher nach Hause?“ ist eine dieser Fragen, die ich mir gelegentlich stelle und die so jedermann nachvollziehen und auf eigene Situationen übertragen kann. Diese Entscheidung ist, obwohl sie eigentlich nur aus zwei Optionen besteht, schon mit einigem Nachdenken verbunden: Entweder kann ich schneller fertig werden mit der Dissertation, wenn ich noch etwas daran arbeite, oder ich kann mich zu Hause entspannen. Ich könnte auch eine Münze werfen und diese darüber entscheiden lassen, was ich nun tun soll und damit jedes Abwägen von Vor- und Nachteilen übergehen. Auch zwei weitere Optionen wären noch denkbar, die diese Entscheidungssituation etwas komplexer gestalten: Ich könnte auch noch zum Sport gehen oder mich mit Freunden in der Stadt treffen. Alle Optionen haben für mich Vor- und Nachteile und ich muss mich irgendwie für eine Option entscheiden.

unbewusste Entscheidungen

Oft erfolgen Entscheidungen auch unbewusst oder *routinisiert*. Dies ist oft bei trivialen Problemen der Fall, bei denen eine tiefergehende und zeitbeanspruchende Beschäftigung sich nicht lohnt: Wenn es draußen stark regnet, wird man sicherlich auf einen abendlichen Spaziergang verzichten und stattdessen lieber zu Hause bleiben. Im Folgenden soll es aber um Entscheidungssituationen gehen, die bewusst erfolgen: Schülerinnen und Schüler

werden explizit dazu aufgefordert, sich gründlich zu überlegen, welche Entscheidung sie treffen wollen und anschließend auch ihren Entscheidungsweg preiszugeben.

1.1.1 Prozesse der Entscheidungsfindung

Rahmenmodell mit 3 Phasen

Aus mehreren Modellen und Beschreibungen haben Betsch und Haberstroh (2005) ein „Rahmenmodell für den Prozess des Entscheidens“ entworfen, das die wesentlichen Aspekte mehrerer Entscheidungstheorien vereint und wichtige Teilprozesse des Entscheidens benennt (siehe Abbildung 1.1). Das Rahmenmodell lässt sich in drei Phasen unterteilen, in die *präselektionale*, die *selektionale* und die *postselektionale Phase*. In der präselektionalen Phase werden Prozesse zusammengefasst, die der eigentlichen Entscheidungsfindung vorausgehen. Dies beinhaltet die Suche nach entscheidungsrelevanten Informationen und das Generieren von möglichen Optionen (Betsch et al. 2011: S. 76). Im Göttinger Modell der Bewertungskompetenz (siehe Kapitel 2.5; Bögeholz 2007, Eggert & Bögeholz 2006) ist diese Phase durch die Teilkompetenz *Generieren und Reflektieren von Sachinformationen* repräsentiert. Die selektionale Phase ist die zentrale Phase einer Entscheidung, während derer die Bewertung der Optionen (sowie deren Konsequenzen) und das Zustandekommen einer Entscheidung vollzogen wird (Betsch et al. 2011: S. 75). Dabei werden die verfügbaren Informationen rekapituliert, bewertet und sich schließlich für eine der Optionen entschieden. Durch die Teilkompetenz *Bewerten, Entscheiden und Reflektieren* wurde auch die selektionale Phase im Göttinger Modell der Bewertungskompetenz aufgegriffen (siehe Kapitel 2.5; Bögeholz 2007, Eggert 2008, Eggert & Bögeholz 2006). In der abschließenden postselektionalen Phase finden u. a. die Anwendung der gewählten Option und die Evaluierung statt. Mit den grauen Pfeilen in Abbildung 1.1 bringen Betsch et al. zugleich zum Ausdruck, dass unsere getroffenen Entscheidungen (bzw. die Konsequenzen daraus) sich immer auch gleichzeitig auf zukünftige Entscheidungssituationen auswirken, indem sie z. B. unser Vorgehen bei ähnlichen Entscheidungen routinisierter gestalten oder von einer bestimmten Option aufgrund schlechter Rückmeldungen aus vorangegangenen Entscheidungen abraten.

1.1.2 Entscheidungsstrategien

Um Entscheidungen zu fällen, gibt es mehrere Möglichkeiten und Vorgehensweisen. Bei der großen Klasse der *analytischen Entscheidungsstrategien*, die beschreiben, dass eine Person eine Bewertung anhand der Konsequenzen einer Option vornimmt, lassen sich Entscheidungsstrategien grundsätzlich in zwei weitere Unterklassen einteilen (Betsch et al.

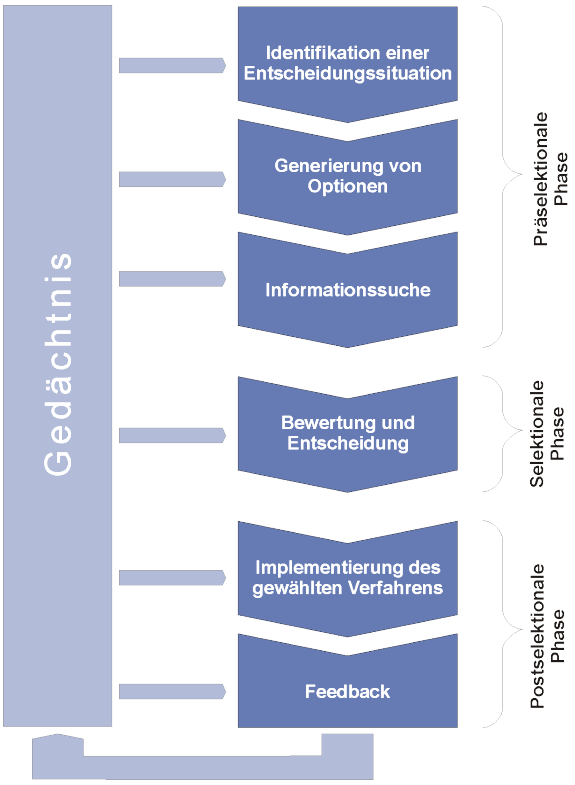

Abbildung 1.1: Rahmenmodell für den Prozess des Entscheidens (nach Betsch et al. 2011, Betsch & Haberstroh 2005: S. 75).

2011: S. 97): Bei den *kompensatorischen Strategien* werden Optionen in all ihren Ausprägungen auf den einzelnen Kriterien zunächst als Ganzes betrachtet und miteinander verglichen. Dadurch ist es möglich, dass schlechte Ausprägungen auf dem einen Kriterium durch bessere Ausprägungen auf einem anderen Kriterium kompensiert werden. Erst dann werden die konkurrierenden Optionen hinsichtlich ihrer Konsequenzen bzw. Ausprägungen geprüft und abschließend die Option gewählt, die dem Entscheider z. B. den größtmöglichen Nutzen verspricht (Betsch et al. 2011: S. 97). Im Gegensatz dazu erfolgt eine solche ganzheitliche Betrachtungsweise nicht bei der Verwendung einer *non-kompensatorischen Strategie*. Dadurch, dass die Ausprägungen von Optionen nur auf einzelnen Kriterien miteinander verglichen werden, ist ein Ausgleich/Kompensation von schlechten Ausprägungen auf einem Kriterium durch eine bessere Ausprägung auf einem anderen Kriterium nicht ohne weiteres möglich: Optionen werden demzufolge nicht in ihrer Gänze betrachtet. Optionen werden also nicht gewählt, wenn ein (oder mehrere) Ausprägungen bestimmte Schwellenwerte nicht erreichen bzw. überschreiten – egal, wie die Ausprägungen der Option bei den anderen Attributen sind (Jungermann et al. 2005: S. 120).

kompensatorisch

non-kompensatorisch

Abseits dieser analytischen Strategien gibt es auch noch die Klasse der *nichtanalytischen*

Tabelle 1.1: Übersicht über einige verschiedene Entscheidungsstrategien (nach Betsch et al. 2011: S. 102). Weitere Entscheidungsstrategien auch in: Jungermann et al. (2005: S. 120-132).

Verschiedene Entscheidungsstrategien (Auswahl)			
Name	Suchmuster zentriert auf	Strategie-klasse	Kurzbeschreibung
Weighted Additive Rule (WADD)[a]	Optionen	komp.	Wähle die Option mit dem höchsten gewichteten Gesamtgewicht.
Equal Weight Rule (EQW)[b]	Optionen	komp.	Wähle die Option mit dem höchsten Gesamtgewicht (gleichgewichtet).
Satisficing Rule (SAT)[c]	Optionen	non-komp.	Wähle die erstbeste Option, die das Anspruchsniveau für jedes Attribut erfüllt.
Elimination by Aspects (EBA)[d]	Attribute	non-komp.	Eliminiere attributsweise die Optionen, die den Kriteriumswert verfehlen. Wähle die Option, die übrig bleibt.
Lexicographic Rule (LEX)[e]	Attribute	non-komp.	Wähle die Option mit dem höchsten Wert auf dem wichtigsten Attribut.

[a] Payne et al. (1988)
[b] Einhorn und Hogarth (1975), Thorngate (1980)
[c] Simon (1955)
[d] Tversky (1972)
[e] Fishburn (1974), Goldstein und Gigerenzer (2002)

Entscheidungsstrategien, die sich z. B. über die Erfahrung aus vorangegangenen Situationen in der Routinisiertheit einer Entscheidung im Alltag oder dadurch äußert, dass wir bereits vorgefasste Meinungen zu Optionen haben (z. B. durch Werbung bei Kaufentscheidungen). Mögliche nichtanalytische Strategien sind aber auch, einfach einem Expertenurteil (z. B. externen Produkttests wie denen der „Stiftung Warentest“) zu folgen oder den Zufall entscheiden zu lassen (Betsch et al. 2011: S. 100). Daran anschließend ist auch ein weiteres Entscheidungsverhalten zu erwähnen, das aus der Erfahrung vorausgehender Studien (Eggert 2008, Eggert & Bögeholz 2010) im Zusammenhang mit Bewertungskompetenz zu beobachten ist: Intuitive Entscheidungen, die im Anschluss gerechtfertigt werden, z. B. durch die Nennung einzelner, passender Kriterien der gewählten Option. Die Optionen und deren Ausprägungen werden hierbei nicht vergleichend betrachtet und gegeneinander abgewogen. Haidt (2001) hat die These aufgestellt, dass alle Entscheidungen zunächst intuitiv getroffen werden. Erst post-hoc werden Begründungen für die getroffene Entscheidung gesucht und zur Rechtfertigung herangezogen. Dabei kommen nur solche Argumente zum Tragen, die die eigene Entscheidung in einem günstigen Licht dastehen lassen, unpassende Kriterien werden in geringerem Umfang oder gar nicht beachtet. Da dieses Vorgehen im Rahmen dieser Studie nicht als Entscheidungs*strategie* im engeren Sinn sondern mangels eines strategischen Vorgehens nur als ein Entscheidungs*verhalten* betrachtet wird, ist es in Tabelle 1.1 nicht mit aufgeführt. Es spielt aber dennoch eine beachtliche Rolle, auch

intuitiv-rechtfertigend

bei Bewertungs- und Entscheidungsaufgaben im Kontext nachhaltiger Entwicklung (siehe Kapitel 2.5.1).

Intuition

Die Frage, welche Auswirkungen solche intuitiven Entscheidungsprozesse beim Bewerten und Entscheiden spielen, ist Gegenstand aktueller Forschung. Menthe (2012), Menthe und Düker (2013) verweisen darauf, dass diesem Aspekt bisher noch zu wenig Aufmerksamkeit geschenkt wird. Wichtig ist, die Rolle der Intuition weiter zu erforschen und daraus sich ergebende etwaige Hemmnisse für das Aneignen adäquater Entscheidungsstrategien zu verstehen (vgl. Menthe 2012). Der Anspruch von Seiten der Fachdidaktiken und der Bildungspolitik an mündige, reflektiert entscheidende und dabei naturwissenschaftliche Erkenntnisse nutzende Bürgerinnen und Bürger in komplexen Situationen nachhaltiger Entwicklung bleibt davon jedoch unberührt (vgl. Bögeholz 2006, Eggert & Hößle 2006).

implizite und explizite Bewertungen

Eine Unterscheidung wird auch zwischen impliziten und expliziten Bewertungen gemacht: Sie können *implizit* vorgenommen werden, wenn der Entscheider sich zwar bewusst ist, dass die zu beurteilenden Optionen mehrere Attribute haben, diese „Bewertungen des Objekts auf den einzelnen Attributen [allerdings] kognitiv ‚irgendwie' integriert [werden]" (Jungermann et al. 2005: S. 120). Eine Option wird somit als „gut" bewertet, obwohl nicht detailliert erkennbar ist, ob der Entscheider diese Option auf allen Attributen als gut betrachtet oder ob diese (Gesamt-) Bewertung schon eine Zusammenfassung von Attributen ist, von denen manche als sehr gut und andere als befriedigend eingestuft werden (Jungermann et al. 2005: S. 120). *Explizit* ist eine Bewertung hingegen, wenn Einzelbewertungen auf jedem Attribut stattfinden, die danach gegebenenfalls zu einer Gesamtbewertung zusammengefasst werden. Dazu zählen z. B. die in Tabelle 1.1 aufgeführten Entscheidungsstrategien.

Im Zusammenhang mit dem Göttinger Modell der Bewertungskompetenz nennt Eggert (2008: S. 30) Strategien, die bei der Operationalisierung der Teilkompetenz *Bewerten, Entscheiden und Reflektieren* einbezogen werden sollten. Hierzu zählt auch das zuletzt genannte intuitive Entscheidungsverhalten, aber in größerem Umfang noch die elaborierteren nonkompensatorischen und kompensatorischen Entscheidungsstrategien.

1.1.3 Kognitiver Aufwand von Schwierigkeiten bei Entscheidungen

Für die Beurteilung von eingesetzten Entscheidungsstrategien können diese nach ihrem kognitivem Aufwand geordnet werden. Alle kognitiven Prozesse können nach einem Ansatz von Newell und Simon (1972) in elementare Teilprozesse zerlegt werden, die Anzahl dieser dient dann wiederum als Maß für den kognitiven Aufwand einer Entscheidungsstrategie (Betsch et al. 2011: S. 102f). Grundsätzlich kann man sagen, dass dieser bei analytischen

Strategien höher ist als bei nichtanalytischen, aber er lässt sich mit der Analyse folgender *elementaren Informationsprozesse* (Betsch et al. 2011, Payne et al. 1993) bestimmen, zu denen u. a. gehören:

elementare Informations-prozesse

- Enkodieren des Wertes einer Konsequenz
- Gewichten eines Wertes
- Addieren von Werten
- Vergleichen eines Wertes mit einem Kriterium
- Eliminieren einer Option
- Beenden des Suchprozesses und die Entscheidung selbst

Zu bemerken ist, dass ein höherer kognitiver Aufwand nicht unbedingt ein Garant für gute Entscheidungen ist. Bestimmte Bedingungen können in Einzelfällen auch dazu führen, dass selbst aufwandsarme Strategien zu ggf. besseren Entscheidungen führen (Payne et al. 1988).

Komplexität

Die Anzahl der Optionen, Attribute der einzelnen Optionen und die Ähnlichkeit der Optionen beeinflussen die Komplexität einer Entscheidungssituation (vgl. Jungermann et al. 2005). Dies hat sich grundsätzlich auch im Kontext physikalischer Leistungstestaufgaben gezeigt: Mit einigen Einschränkungen über die erwartete Abfolge postulierter Kompetenzniveaus konnte Kauertz (2007: S. 122) zumindest einen Zusammenhang zwischen der Anzahl von Inhaltselementen und der Aufgabenschwierigkeit feststellen. Dies wird sich auch beim Projekt „Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I (ESNaS)“ zunutze gemacht, in dem die Komplexität von Aufgaben gezielt variiert wird (siehe Kapitel 2.2.3, Hostenbach 2011: S. 40ff).

1.1.4 Entscheiden unter Risiko

Technikfolgen-abschätzung

Das Entscheiden unter Risiko oder Unwissenheit wird hier im Sinne einer Folgenabschätzung von neuartigen Technologien verstanden, wie es dem Standard B3 der KMK-Bildungsstandards für Physik entspricht (siehe Tabelle 2.1, KMK 2005c). Es geht dabei um eine Technikfolgenabschätzung und die Frage, welche Chancen und Risiken sich für die Gesellschaft aus der Anwendung einer neuen Technologie ergeben.

Sensibilität für Risiken

Risiken angemessen abzuschätzen ist allgemein ein komplexes Unterfangen, gerade wenn es um „große“ technologische Neuerungen geht, sind sich selbst Experten uneinig, wie die Beispiele Gentechnologie, Nanotechnologie und Kernenergie einleuchtend demonstrieren.

Beck (2007: S. 211ff) schreibt in diesem Zusammenhang von einem „Nicht-Wissen-Können“ über Risiken und Unsicherheiten, das in Zukunft noch weiter zunehme. Übertriebene Sicherheitsversprechen der Wissenschaft und Technik tragen zu einem Vertrauensverlust der Gesellschaft bei, wie z. B. Unfälle bei Kernenergieanlagen in Tschernobyl, Harrisburg (vgl. Bonß 1996: S. 169ff) oder Fukushima zeigen. Das Wahrnehmen, Einschätzenkönnen und Kommunizieren von Risiken ist für einen kompetenten Umgang mit Entscheidungs- bzw. Gestaltungsaufgaben im Kontext nachhaltiger Entwicklung von zentraler Bedeutung (Bögeholz & Barkmann 2005, de Haan et al. 2008).

1.2 Nachhaltigkeit und nachhaltige Entwicklung

Historischer Begriff der Forstwirtschaft

Die Genese des Worts *Nachhaltigkeit* führen de Haan et al. (2008: S. 50ff) auf das Jahr 1713 und den Oberberghauptmann H. C. von Carlowitz zurück, der diesen Begriff als erster geprägt und ihn im Rahmen der Forstwirtschaft etabliert haben soll (vgl. Grober 1999). Noch konkreter aber wird der Begriff von Hartig (1791) gefasst:

> „Es läßt sich keine dauerhafte Forstwirtschaft denken und erwarten, wenn die Holzabgabe aus den Wäldern nicht auf Nachhaltigkeit berechnet ist. Jede weise Forstdirektion muss daher die Waldungen [...] so hoch als möglich, doch so zu benutzen suchen, dass die Nachkommenschaft wenigstens ebensoviel Vorteil daraus ziehen kann, wie sich die jetzt lebende Generation zueignet.“

Parallelen zu heute

Das damalige Problem, dass zum Heizen, Werken, Kochen, etc. mehr Holz geschlagen wurde als in gleicher Zeit nachwachsen kann, lässt sich 1:1 auf heutige Verhältnisse und den Umgang mit Ressourcen aller Art verallgemeinern. Auch für die von ihm vorgeschlagenen Maßnahmen lassen sich heutige Pendants im Kontext regenerativer Energien finden (z. B. „Holzsparkunst“ – Energieeinsparung, Nutzung geschlossener Öfen statt offener – Steigerung der Energieeffizienz, vgl. de Haan et al. 2008: S. 71). Auch der damalige Appell an die Bestandsicherungsinteressen der fürstlichen Herrschaftsfamilien zur Verdeutlichung des Nachhaltigkeitsprinzips (vgl. de Haan et al. 2008: S. 51) lässt sich heutzutage auf das Gemeinwohl und die Zukunftsfähigkeit in einer globalisierten Welt anwenden.

WCED-Leitbild und Retinität

Nach Jahr(zehnt)en des industriellen Wachstums und des steigenden Wohlstands traten jedoch in den 1980er Jahren immer mehr irreversible Umweltschäden zutage (vgl. Büttner 2001). Über den Fortgang dieser Entwicklung besorgt wurde in einem Bericht der *World Commission on Environment and Developement (WCED)* ein Leitbild einer wirtschaftlichen Entwicklung eingeführt, „[...] that meets the needs of the present, without compromising the ability of future generations, to meet their own needs“ (WCED 1987: S. 37). Es knüpft

damit unmittelbar an das eingangs erwähnte Prinzip der Ressourcennachhaltigkeit aus der Forstwirtschaft an, ist aber noch weiter gefasst und bezieht neben der Umweltgerechtigkeit für diese und zukünftige Generationen auch noch (Grund-)Bedürfnisorientierung und das Zusammenspiel von Ökologie, Ökonomie und Sozialem („Retinität“, SRU 1994) mit ein (Bögeholz 2000, WCED 1987).

Heutiges Verständnis

Das weltweite Leitbild einer nachhaltigen Entwicklung wurde in dem 1992 verabschiedeten Aktionsprogramm *Agenda 21* auf der *United Nations Conference on Environment and Development* (UNCED) von 178 Staaten in Rio de Janeiro beschlossen (UNCED 1992). Aus diesem Bestreben heraus riefen die Vereinten Nationen die Weltdekade *Bildung für nachhaltige Entwicklung* aus, die im Jahr 2005 begann und noch bis 2014 andauert. Hiermit verbunden sind eine Vielzahl nationaler und internationaler Bemühungen und Projekte zur Bewerbung und Verankerung des Nachhaltigkeitsprinzips im Denken und Handeln der Menschen (vgl. UNESCO 2005).

1.3 Energiebedarf und Erneuerbare Energiequellen

Anstieg des Energiebedarfs

Der weltweite Energiebedarf wird nach aktuellen Schätzungen im Jahr 2035 etwa 73% über dem des Jahres 2010 liegen, begrenzt auf die OECD-Staaten wird in diesem Zeitraum ein Zuwachs von 24% erwartet (IEA 2012: S. 180). Wenngleich fossile Energieträger auch weiterhin wichtig für die Aufrechterhaltung der Energieversorgung sind und ihr Anteil weiterhin anwachsen wird, so kommt den Erneuerbaren Energieträgern im Hinblick auf eine nachhaltige Energiewirtschaft eine immer größer werdende Bedeutung zu: Verglichen mit 2012 wird ihr Anteil bis 2035 weltweit um voraussichtlich 170% zunehmen (siehe Tabelle 1.2). Unter Erneuerbaren Energieträgern werden Photovoltaik, Biomasse, Wasserkraft und besonders Windkraft verstanden (IEA 2012: S. 211), die gerade in Norddeutschland sehr präsent und damit auch Schülerinnen und Schülern vertraut sind.

20-20-20-Ziel und Energiewende

Innerhalb der Europäischen Union hat man sich 2007 auf das *20-20-20-Ziel* verständigt, der Reduktion von Treibhausgasen um 20% unter das Level von 1990, dem Ausbau regenerativer Energien auf 20% des Energieverbrauchs und einer 20%igen Energieeinsparung (EC 2007a). Der Windenergie kommt dabei die größte Bedeutung zu, ihr Anteil allein soll 12% des EU-Strombedarfs betragen (davon ein Drittel aus Offshore-Windenergieanlagen, EC 2007b: S. 11). Für das Jahr 2050 hat sich die Europäische Kommission das Ziel eines Anteils von 50% an Erneuerbaren Energien vorgenommen (EC 2011: S. 7). Die Forderung nach vermehrter Nutzung erneuerbarer Energiequellen und Deutschlands Bemühungen zum Einleiten der „Energiewende“ ist auch in den internationalen Medien präsent, „Energiewende“ hat es als Begriff in englischsprachige Zeitungen geschafft (Dempsey 2012).

Tabelle 1.2: Schätzungen des Energiebedarfs und Anteil verschiedener Energieträger an der Stromversorgung (nach IEA 2012: S. 180,182). Grundlage: New-Policies-Szenario der IEA (IEA 2012: S. 34f). Absolutwerte in TWh, Prozentangaben bezogen auf 2010.

	Energiebedarf im Jahr...			*Erzeugte Energie nach Energieträger*					
Region:	2010	2035			2010	2020		2035	
OECD-Staaten	9618	11956	+24%		10848	11910	+10%	13297	+23%
				fB:[a]	6600	6629	+0%	6401	-3%
				Ke:[a]	2288	2317	+1%	2460	+8%
				EE:[a]	1960	2963	+51%	4435	+126%
Nicht-OECD-Staaten	8825	19903	+126%		10650	16325	+53%	23340	+119%
				fB:[a]	7847	11163	+42%	14528	+85%
				Ke:[a]	468	1125	+140%	1906	+307%
				EE:[a]	2245	4037	+80%	6905	+208%
Weltweit	18443	31859	+73%		21408	28235	+32%	36637	+71%
				fB:[a]	14446	17793	+23%	20929	+45%
				Ke:[a]	2756	3443	+25%	4366	+58%
				EE:[a]	4206	6999	+66%	11342	+170%

[a] davon aus fossilen Brennstoffen (fB), aus Kernenergie (Ke) oder aus Erneuerbaren Energiequellen (EE).

Speicherbedarf

Im Gegensatz zu den fossilen Brennstoffen und zur Kernenergie, die auch weiterhin eine wichtige Rolle bei der Energieversorgung spielen werden, sind Erneuerbare Energien z. T. von äußeren Bedingungen abhängig. Je nach Intensität von z. B. Sonne und Wind gibt es auch Schwankungen in der Energieerzeugung, die nicht von vornherein planbar sind. Weiterer Forschungsbedarf besteht daher für die Entwicklung kostengünstiger Kurzzeit- und Langzeitenergiespeicher, die überschüssige Energie aufnehmen und sie im Bedarfsfall wieder abgeben können.

Für die Überbrückung von Spitzenlastzeiten (z. B. während der Mittags- und Abendstunden) werden auch heute bereits verschiedene Speichertechnologien eingesetzt. Die bekannteste unter ihnen ist wohl ein Pumpspeicherkraftwerk, in dem nachts überschüssige Energie dazu verwendet wird, Wasser von einem tieferen in ein höher gelegenes Bassin zu pumpen und es dann im Bedarfsfall durch Turbinen zurückfließen zu lassen und somit elektrische Energie zu erzeugen. Zum Ausgleich von Schwankungen bei den regenerativen Energien werden weitere Energiespeicher benötigt, die geeignet sind die Verfügbarkeit Erneuerbarer Energien zu verbessern.

Neben geeigneten Speichertechnologien kommen aber auch eine verbesserte Netzinfrastruktur und die intelligente Vernetzung von Kraftwerken in Frage. Unterschiede im regionalen Windaufkommen können somit durch Windenergieanlagen aus einer anderen Region oder andere Kraftwerksarten kompensiert werden (VDE 2009: S. 36).

KAPITEL 2

Bewertungskompetenz: Definition und Stand der Forschung

2.1 Kompetenzen

Bedeutungen

Der Kompetenzbegriff wird in vielerlei Hinsicht und in unterschiedlichen Domänen verwendet. In Politik und Jura versteht man unter Kompetenz einen Zuständigkeitsbereich, in den Wirtschaftswissenschaften ist von Führungskompetenz der Geschäftsführung oder den Kernkompetenzen eines Unternehmens die Rede, in der Sprachwissenschaft von der Fähigkeit, sich sprachlich zu äußern. Vielen Kompetenzbegriffen gemeinsam ist, dass sie getreu der lateinischen Übersetzung (von *competere* – zu etwas fähig sein) bestimmte Fähigkeiten bezeichnen. In den Sozial- und Erziehungswissenschaften hat der Kompetenzbegriff vor etwa fünfzig Jahren durch die von Noam Chomsky entwickelte Theorie der Sprachkompetenz Einzug gehalten (Chomsky 1968, Klieme & Hartig 2008: ausführlich).

2.1.1 Der Kompetenzbegriff im Kontext von Bildungsstandards

Definition nach Weinert

Aus der Vielzahl der Definitionen von Kompetenzen haben Klieme et al. (2003) die Definition von Weinert (2001) ausgewählt, die in der Bildungsforschung mittlerweile etabliert ist:

> „Kompetenzen sind die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (Weinert 2001: S. 27f).

Er unterscheidet weiterhin *fachliche Kompetenzen*, *fachübergreifende Kompetenzen* und *Handlungskompetenzen*, hebt hervor, „dass diese [...] Kompetenzen für ein gutes und erfolgreiches Leben innerhalb wie außerhalb der Schule notwendig sind“ (Weinert 2001: S. 28) und plädiert anschließend für eine Ausweitung der vergleichenden schulischen Leistungsmessung, die zum damaligen Zeitpunkt noch nicht so etabliert und umstritten war (Brügelmann 2001).

Funktion von Bildungsstandards

Die Kultusministerkonferenz (KMK) fordert aber auch, dass „der Auftrag der schulischen Bildung [...] weit über die funktionalen Ansprüche von Bildungsstandards hinaus [geht]. [...] Schülerinnen und Schüler sollen zu mündigen Bürgerinnen und Bürgern erzogen werden, die verantwortungsvoll, selbstkritisch und konstruktiv ihr berufliches und privates Leben gestalten und am politischen und gesellschaftlichen Leben teilnehmen können“ (KMK 2005d: S. 6).

Nachhaltigkeit in der Bildung

Das Konzept der Nachhaltigkeit wird in den KMK-Bildungsstandards allerdings nicht einheitlich verfolgt, die nationalen Bildungsstandards für Physik verwenden im Gegensatz zu denen der Biologie zudem nicht explizit den Begriff *Nachhaltigkeit* (siehe auch Kapitel 2.2). de Haan et al. (2008: S. 183ff) diskutieren *Bildung für Nachhaltigkeit und Gerechtigkeit* detailliert und geben Anregungen für eine Bildung zu nachhaltiger und gerechter Entwicklung als fächerübergreifendes Ziel schulischer Bildung. Sie gehen dabei vom Konzept für Schlüsselkompetenzen der OECD (2005a) aus, das von anderen Konzepten zu fachlichen Kompetenzen abgegrenzt werden muss und aus übergreifenden Bildungszielen normativ abgeleitet wurde (de Haan et al. 2008: S. 184). In den gegebenen Handlungsempfehlungen finden sich neben der generellen Forderung nach einer stärkeren Betonung einer auf Nachhaltigkeit ausgerichteten Schulbildung auch Aspekte wieder, die direkten Bezug zu der Intention dieser Studie haben: Mit dem *rationalen Bewerten gegenwärtiger und zukünftiger Handlungswirkungen* und der *angemessenen Wahrnehmung und Bewältigung von Risiken* werden zwei zentrale Ideen angesprochen, die sich auch in dem hier vorgestellten Testinstrument widerspiegeln (de Haan et al. 2008: S. 233).

2.1.2 Entstehung der KMK-Bildungsstandards

„PISA-Schock“

Die Leistungsfähigkeit des deutschen Bildungssystems wurde in den 1990er Jahren in der öffentlichen Meinung überschätzt. Als deutsche Schülerinnen und Schüler bei großen internationalen Vergleichsstudien, wie z. B. der TIMSS oder der PISA-Studie der OECD (Baumert et al. 2001), nur mittelmäßig abschnitten und Deutschland in den drei Kompetenzbereichen Mathematik, Lesefähigkeit und Naturwissenschaft nur unter dem OECD-Durchschnitt rangierte, war der Aufschrei groß und die Politik einem Handlungsdruck ausgesetzt (Schott

& Azizi Ghanbari 2012: S. 16ff). Dem Vorbild erfolgreicherer Staaten folgend, versprach die Einführung von Bildungsstandards hier Abhilfe: Eine richtungsweisende und vom Bundesministerium für Bildung und Forschung in Auftrag gegebene Studie führender Bildungswissenschaftler (Klieme et al. 2003) war letztlich das Fundament für den Entschluss der KMK im Jahr 2004, in den Fächern Deutsch, Erste Fremdsprache, Mathematik und Naturwissenschaften die nationalen Bildungsstandards für das Ende der Sekundarstufe I einzuführen. Ziel ist es seitdem festzuschreiben, was (welche Kompetenzen) Schülerinnen und Schüler am Ende von Klasse 10 können sollten (was also am Ende herauskommt, „Output-Steuerung“) und nicht mehr nur vorzuschreiben, welche Inhalte und Themen bis dahin im Unterricht behandelt werden sollen („Input-Steuerung“) (Schott & Azizi Ghanbari 2012).

Regel- statt Mindeststandards

Klieme et al. (2003: S. 27) empfahlen „jedoch nachdrücklich [...], in den nationalen Bildungsstandards für Deutschland ein verbindliches Minimalniveau festzuschreiben. [...] Diese Konzentration auf Mindeststandards ist für die Qualitätssicherung im Bildungswesen von entscheidender Bedeutung.“ Von dieser eindeutigen Empfehlung zur Festschreibung dessen, was Schülerinnen und Schüler mindestens am Ende von Klassse 10 können sollten, ist die KMK abgewichen und hat stattdessen Regelstandards beschlossen, die angeben, was Schülerinnen und Schüler durchschnittlich können sollten (Schott & Azizi Ghanbari 2012: S. 22f). Sie begründet dies unter anderem damit, dass „notwendige Mindeststandards erst nach einem längeren Prozess der Erfahrung im Umgang mit Bildungsstandards

Abbildung 2.1: Der Titel der Zeitschrift „Der Spiegel“ vom 10.12.2001 drückt das Erstaunen aus, das mit dem mäßigen Abschneiden deutscher Schülerinnen und Schüler bei PISA einher ging.

formuliert werden können“, „Niveaustufen [bereits] präzisiert und insgesamt die Standards auf Aufgabenbeispiele validiert“ sein müssen und verweisen auf die bestehende Gefahr, „Schülerinnen und Schüler massiv zu unterfordern oder [...] durch überzogene Bildungsstandards zu überfordern“, solange diese Aspekte nicht ausreichend erforscht wurden (KMK 2005d: S. 14). Sie stellt weiter fest: „Die Standards müssen zukünftig validiert werden“ (KMK 2005d: S. 14).

fehlende fachdidaktische Forschung

Zum Zeitpunkt der Einführung 2005 gab es hinsichtlich der Kompetenzen nur wenig fachdidaktische Forschung. Deren Evaluierung und die Entwicklung von Messinstrumenten konnte und kann erst im Nachhinein erfolgen, u. a. im Rahmen des Projekts *Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I (ESNaS)* (siehe Kapitel 2.2.3). Das hier in dieser Studie vorgestellte Testinstrument kann somit einen Beitrag zur Messung von Bewertungskompetenz leisten.

2.1.3 Umsetzung in den Bundesländern

Bildung ist Ländersache

Zwar haben sich die Bundesländer zur Umsetzung der nationalen Bildungsstandards verpflichtet, jedes Bundesland kann diese jedoch noch ausdifferenzieren und ergänzen. Dabei wird in den einzelnen Bundesländern unterschiedlich vorgegangen: In Niedersachsen gibt es ein *Kerncurriculum*, das für die Naturwissenschaften in der Sekundarstufe I seit 2007 (NdsMK 2007) und für Physik in der gymnasialen Oberstufe seit 2010 gilt (NdsMK 2009). Für jeden der Kompetenzbereiche *Fachwissen*, *Erkenntnisgewinnung*, *Kommunikation* und *Bewertung* gibt es somit detaillierte Vorgaben, welche Kompetenzen Schülerinnen und Schüler am Ende der Klassenstufen 6, 8, 10 und 12 erreichen sollten.

Tabelle 2.1: Bewertungkompetenzen in den naturwissenschaftlichen Unterrichtsfächern (aus KMK 2005a, b, c).

Biologie		*Chemie*		*Physik*	
Die Schülerinnen und Schüler ...					
B 1	unterscheiden zwischen beschreibenden (naturwissenschaftlichen) und normativen (ethischen) Aussagen,	B 1	stellen Anwendungsbereiche und Berufsfelder dar, in denen chemische Kenntnisse bedeutsam sind,	B 1	zeigen an einfachen Beispielen die Chancen und Grenzen physikalischer Sichtweisen bei inner- und außerfachlichen Kontexten auf,
B 2	beurteilen verschiedene Maßnahmen und Verhaltensweisen zur Erhaltung der eigenen Gesundheit und zur sozialen Verantwortung,	B 2	erkennen Fragestellungen, die einen engen Bezug zu anderen Unterrichtsfächern aufweisen und zeigen diese Bezüge auf,	B 2	vergleichen und bewerten alternative technische Lösungen auch unter Berücksichtigung physikalischer, ökonomischer, sozialer und ökologischer Aspekte,
B 3	beschreiben und beurteilen Erkenntnisse und Methoden in ausgewählten aktuellen Bezügen wie zu Medizin, Biotechnik und Gentechnik, und zwar unter Berücksichtigung gesellschaftlich verhandelbarer Werte,	B 3	nutzen fachtypische und vernetzte Kenntnisse und Fertigkeiten, um lebenspraktisch bedeutsame Zusammenhänge zu erschließen,	B 3	nutzen physikalisches Wissen zum Bewerten von Risiken und Sicherheitsmaßnahmen bei Experimenten, im Alltag und bei modernen Technologien,
B 4	beschreiben und beurteilen die Haltung von Heim- und Nutztieren,	B 4	entwickeln aktuelle, lebensweltbezogene Fragestellungen, die unter Nutzung fachwissenschaftlicher Erkenntnisse der Chemie beantwortet werden können,	B 4	benennen Auswirkungen physikalischer Erkenntnisse in historischen und gesellschaftlichen Zusammenhängen.
B 5	beschreiben und beurteilen die Auswirkungen menschlicher Eingriffe in einem Ökosystem,	B 5	diskutieren und bewerten gesellschaftsrelevante Aussagen aus unterschiedlichen Perspektiven,		
B 6	bewerten die Beeinflussung globaler Kreisläufe und Stoffströme unter dem Aspekt der nachhaltigen Entwicklung,	B 6	binden chemische Sachverhalte in Problemzusammenhänge ein, entwickeln Lösungsstrategien und wenden diese an.		
B 7	erörtern Handlungsoptionen einer umwelt- und naturverträglichen Teilhabe im Sinne der Nachhaltigkeit.				

2.2 Bewertungskompetenz in den naturwissenschaftlichen Unterrichtsfächern – Gemeinsamkeiten und Unterschiede

Biologie, Chemie und Physik

Bewertungskompetenzen sind zwar für alle naturwissenschaftlichen Fächer definiert, werden in den einzelnen Fächern aber nicht einheitlich verstanden. In der Biologie werden in diesem Zusammenhang vorrangig Fragestellungen zur Gestaltung einer nachhaltigen Entwicklung und globale Kreisläufe genannt (siehe Tabelle 2.1). Diesem Verständnis am ehesten entspricht in der Physik z. B. das Vergleichen und Bewerten unter Berücksichtigung physikalischer, ökonomischer und ökologischer Aspekte, aber darüberhinaus werden u. a. auch das Nutzen physikalischen Wissens bei Sicherheitsmaßnahmen beim Experimentieren, im Alltag und bei modernen Technologien als Bewertungskompetenz aufgefasst (KMK 2005c) und somit eine Komponente beschrieben, die sich als *Umgang mit Risiken* bezeichnen ließe. Die Bildungsstandards aller naturwissenschaftlichen Fächer werden hierzu noch konkreter und erinnern daran, dass die „naturwissenschaftlich-technische Entwicklung auch Risiken [birgt], die erkannt, bewertet und beherrscht werden müssen. Hierzu ist Wissen aus den naturwissenschaftlichen Fächern nötig. Naturwissenschaftliche Bildung ermöglicht dem Individuum eine aktive Teilhabe an gesellschaftlicher Kommunikation und Meinungsbildung über technische Entwicklungen und naturwissenschaftliche Forschung und ist deshalb wesentlicher Bestandteil von Allgemeinbildung“ (KMK 2005a, b, c: jeweils S. 6).

Nachhaltigkeit

Der Aspekt der Nachhaltigkeit wird in den Bildungsstandards besonders im Fach Biologie berücksichtigt (B6, B7, siehe Tabelle 2.1). Bei den Bildungsstandards Chemie taucht unter Bewertungskompetenz kein Nachhaltigkeitsaspekt auf, im Fließtext zum Beitrag des Faches Chemie zur Bildung findet sich jedoch die Formulierung, dass Schülerinnen und Schüler „gleichzeitig [...] für eine nachhaltige Nutzung von Ressourcen sensibilisiert [werden sollen]“ (KMK 2005b). In den Bildungsstandards der Physik hingegen taucht der Begriff der Nachhaltigkeit gar nicht auf (KMK 2005c), allerdings werden mit dem Standard B2 ökologische, ökonomische und soziale Aspekte aufgegriffen, die zentrale Elemente des Leitbildes einer nachhaltigen Entwicklung darstellen (siehe Kapitel 1.2; Bögeholz 2011).

Erdkunde

Für das Fach Erdkunde hat die Deutsche Gesellschaft für Geographie analog zur KMK eigene Bildungsstandards entwickelt, die sechs Kompetenzbereiche unterscheiden, von denen einer den Titel „Beurteilen und Bewerten“ trägt (DGFG 2012). Den Bundesländern ist freigestellt, auch diese Bildungsstandards in ländereigene Curricula zu übernehmen, in Niedersachsen gibt es z. B. seit 2011 ein Kerncurriculum für das Fach Erdkunde, in dem fünf verschiedene Kompetenzbereiche (u. a. „Beurteilen und Bewerten“ unterschieden werden (NdsMK 2010). Aspekte wie das „Bewerten [des] Entscheidungsprozesses [...]“ oder

das „Bewerten [des] Aussagewerts verwendeter Materialien“ werden dabei ebenfalls dem Kompetenzbereich „Beurteilen und Bewerten“ zugeschrieben.

2.2.1 Vorläufer von Bewertungskompetenz

früher kein eigenständiger Begriff

Bewertungskompetenz ist seit der Verabschiedung der Bildungsstandards durch die Kultusministerkonferenz im Jahr 2004 zu einem populären und aktuellen Forschungsthema geworden. Darauf, dass sie allerdings nicht etwas gänzlich Neues ist, verweisen Knittel und Mikelskis-Seifert (2011a), wenngleich damalige Veröffentlichungen eine andere Bezeichnung verwendeten. Beispielhaft sei hier auf die Dissertationsschrift von Mikelskis (1979) verwiesen: Bereits der Titel „Zum Verhältnis von Wissenschaft und Lebenswelt im Physikunterricht – dargestellt am Thema Kernkraftwerke“ macht deutlich, dass es in dieser Schrift um einen Kernbereich von Bewertungskompetenz geht, der Schnittstelle zwischen Naturwissenschaften und Gesellschaft.

Aber auch Duit (1986) beschreibt in seiner Habilitationsschrift die gesellschaftlichen Aspekte des Energiebegriffs, indem er darauf verweist, dass „physikalische Bildung zu verantwortungsbewusstem gesellschaftspolitischen Handeln [beiträgt und insbesondere] eine sachbezogene öffentliche Diskussion physikalischer Technologien, ihrer Vorzüge, aber auch ihrer Probleme [ermöglicht]“ (Duit 1986: S. 102) und damit auch Aspekte dessen aufgreift, was wir heutzutage unter Bewertungskompetenz verstehen.

Bewertungskompetenz stellt also nicht etwas gänzlich Neues dar, ist aber in seiner Konkretisierung durch die Bildungsstandards und Kerncurricula der Länder ein relativ neuartiges Konstrukt, für das es aktuell noch Forschungsbedarf gibt und das bei Lehrerinnen und Lehrern auch noch nicht einheitlich verstanden wird[1].

2.2.2 Überblick über Interpretationen von Bewertungskompetenz

KMK

Der Begriff Bewertungskompetenz beschreibt, inwiefern Schülerinnen und Schüler in der Lage sind, naturwissenschaftliche Sachverhalte in verschiedenen Kontexten zu erkennen und zu bewerten und ist somit eine Grundvoraussetzung dafür, sich an gesellschaftlichen Diskursen zu beteiligen, verschiedene Perspektiven einzunehmen, dabei gesellschaftlich

1 Einen Beitrag zur „Verunsicherung“ von Lehrerinnen und Lehrern stellt sicherlich dar, dass die Bildungsstandards in kürzester Zeit erarbeitet wurden, ohne dass es bis dahin ausreichende fachdidaktische Forschung gab (Sill 2004) und die Lehrerinnen und Lehrer deshalb mit der „Interpretation“ allein gelassen wurden. Schecker und Höttecke (2007) weisen zudem darauf hin, dass zumindest die Beispielaufgaben der Bewertungskompetenz in Physik unzureichend bzw. schlecht gewählt und nicht eindeutig den Kompetenzbereichen zuordenbar sind.

verhandelte/gesetzte Werte, Normen und Grenzen zu berücksichtigen und Entscheidungen sachgerecht und verantwortungsbewusst zu treffen (KMK 2005a, b, c). Im Folgenden werden einige Sichtweisen auf Bewertungskompetenz vorgestellt, die im weiteren Verlauf des Kapitels konkretisiert werden und verdeutlichen, unter welchem Aspekt Bewertungskompetenz in dieser Arbeit verstanden wird.

In der Biologie haben sich im Wesentlichen zwei Definitionen von Bewertungskompetenz durchsetzen können: Die in dieser Arbeit verwendete Definition von Bögeholz (2007: S. 209) fasst Bewertungskompetenz als die „Fähigkeit, sich in komplexen Problemsituationen begründet und systematisch bei unterschiedlichen Handlungsoptionen zu entscheiden, um kompetent am gesellschaftlichen Diskurs um die Gestaltung von nachhaltiger Entwicklung teilhaben zu können“. Diese Definition legt somit einen Schwerpunkt auf nachhaltige Entwicklung. Ethische Aspekte und der Einfluss von gesellschaftlichen Werten und Normen, wie sie im Biologieunterricht zum Beispiel im Kontext der Gentechnologie von Wichtigkeit sind, werden dabei jedoch auch berücksichtigt. Definitionen, die zum *Modell der ethischen Urteilskompetenz* passen, beschreiben Bewertungskompetenz als die Fähigkeit „ethische Relevanz naturwissenschaftlicher Themen wahrzunehmen, damit verbundene Werte zu erkennen und abzuwägen sowie ein reflektiertes und begründetes Urteil zu fällen“ (Mittelsten Scheid & Hößle 2008: S. 88).

Nachhaltigkeit

Ethik

Eine Verknüpfung dieser beiden Schwerpunkte einerseits mit dem individuellen Gewichten von Bewertungskriterien andererseits unternimmt Rost (2002), indem er von Schülerinnen und Schülern fordert, „Entscheidungs- und Handlungsalternativen gegeneinander abzuwägen, sich dabei der involvierten Wertvorstellungen bewusst zu werden und hypothetische oder tatsächliche Entscheidungen aufgrund einer persönlichen Gewichtung vorzunehmen“ um Bewertungskompetenz bei Schülerinnen und Schülern zu fördern.

Gewichten von Kriterien

Wiederum für das Fach Biologie wird Bewertungskompetenz von Mayer et al. (2004) charakterisiert als „Fahigkeit, Kriterien heranzuziehen und zu gewichten, um deskriptives Wissen über den zu beurteilenden Sachverhalt mit individuellen oder gesellschaftlichen Wertsetzungen in transparenter Weise zu verknüpfen“ (Mayer et al. 2004).

Kriterien heranziehen

Mayer et al. (2004) postulieren Niveaus von Bewertungskompetenz: „Kompetenzstufen ergeben sich aus der Unterscheidung zwischen dem Anwendenkönnen vorgegebener Bewertungskriterien bzw. der Arbeit mit selbstentwickelten Bewertungskriterien [...]. Höhere Stufen des Erwerbs von Bewertungskompetenz sind durch zusätzliche Fähigkeiten charakterisiert, z. B. dass Lernende den Bewertungsprozess an sich sowie die Bewertungsproblematik [...] reflektieren sowie Bewertungskriterien hinsichtlich übergeordneter Wertorientierungen systematisieren können“.

Niveaus

Physik: meist innerfachliches Bewerten

Bewertungskompetenz ist im Fach Biologie am weitesten ausdifferenziert, stellt im Vergleich zu Chemie und Physik am meisten Anknüpfungspunkte an die Gesellschaft her (siehe Tabelle 2.1) und weist am meisten Schnittpunkte mit Socio-Scientific Issues (SSI, siehe Kapitel 2.3) auf. Demgegenüber verweisen Schecker und Höttecke (2007) darauf und kritisieren gleichzeitig, dass die beschriebenen Standards für Bewertungskompetenz hauptsächlich nur auf innerfachliches Bewerten abzielen. Eine Ausnahme wird hier sicherlich der Standard B2 (siehe Tabelle 2.1) darstellen, auf den im Rahmen dieser Arbeit und der entwickelten Aufgaben besonders Bezug genommen werden kann.

physikbezogene Projekte

Untersuchungen zur Bewertungskompetenz im Physikunterricht sind zurzeit noch sehr rar und beziehen sich auf einzelne Ideen und Aufgabenbeispiele, wie z. B. Höttecke und Mrochen (2010) und Schecker und Höttecke (2007), die u. a. auch die Standortwahl einer Windenergieanlage diskutieren. Knittel und Mikelskis-Seifert (2011b, 2012, 2013) beschreiben eine Unterrichtseinheit (Interventionsstudie) zum Thema Photovoltaik, die auch auf dem Göttinger Modell der Bewertungskompetenz basiert.

Als weiteres Beispiel für die Charakterisierung von Bewertungskompetenz im Physikunterricht ist das Projekt *Der Klimawandel vor Gericht* zu nennen, das über das Fach Physik hinaus auch Ansätze für eine fächerübergreifende Definition von Bewertungskompetenz liefert und mit vielen Unterrichtsmaterialien, Rollen- und Planspielen rund um das Oberthema *Klimawandel* angereichert ist (Eilks et al. 2011). Bewertungskompetenz wird hier beschrieben als „Voraussetzung für Mündigkeit und ermöglicht die reflektierte Teilhabe an gesellschaftlichen Kontroversen und Entscheidungen“ (Eilks et al. 2011: S. 9f).

fehlende Testinstrumente

Es bleibt festzustellen, dass für die Physik derzeit zwar Einzelstudien und Materialien vorliegen um „irgendeinen Aspekt von Bewertungskompetenz“ mit Schülerinnen und Schülern zu behandeln, groß angelegte Studien, die Bewertungskompetenz in einen größeren Zusammenhang rücken, sind bisher jedoch noch nicht durchgeführt worden. Es fehlt zudem bisher an geeigneten und evaluierten Testinstrumenten – diese Dissertation soll daher einen Beitrag zur Schließung dieser Lücke sein. Dem Thema *Bewertungskompetenz im Physikunterricht* wird sich auch im Rahmen des Projekts *Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I* (ESNaS) verstärkt gewidmet.

2.2.3 Das ESNaS-Modell für Bewertungskompetenz

Bewertungskompetenz im Jahr 2018

In dem Projekt „Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I“ (ESNaS) werden seit dem Jahr 2012 erstmals die 2004 beschlossenen Nationalen Bildungsstandards in einer groß angelegten Studie evaluiert. Dabei waren 2012 die Kompetenzbereiche *Fachwissen* und *Erkenntnisgewinnung* Bestandteile der Erhebung, im Jahr

2018 werden in einem weiteren Schritt die Kompetenzen der Schülerinnen und Schüler in den Bereichen *Kommunikation* und *Bewertung* untersucht (Hostenbach 2011: S. 8). Um möglichst differenzierte Aussagen treffen zu können, soll im Projekt ESNaS ein Kompetenzmodell für Bewertung Anwendung finden, das Bewertungskompetenz möglichst unabhängig von den anderen Kompetenzbereichen erfassen und darstellen kann (Hostenbach 2011: S. 40). Dabei konnte auf vorangegangene Studien und Kompetenzmodelle zu den Kompetenzbereichen *Fachwissen* und *Erkenntnisgewinnung* zurückgegriffen werden (u. a. Kauertz et al. 2012, Mannel 2010, Ropohl 2010, Walpuski et al. 2008, 2010, Walpuski & Sumfleth 2010).

dreidimensionales Modell

Im ESNaS-Kompetenzstrukturmodell wird Bewertungskompetenz in einem mehrdimensionalen Modell abgebildet. Somit sind im ESNaS-(Gesamt-)Modell alle Kompetenzbereiche und die Dimensionen *Kognitive Prozesse* und *Komplexität* enthalten. Die Teilbereiche zu jedem der vier Kompetenzbereiche *Erkenntnisgewinnung*, *Fachwissen*, *Kommunikation* und *Bewertung* sind im Unterschied zu den anderen beiden Dimensionen nicht hierarchisch gegliedert. Im Rahmen der Bewertungskompetenz werden *Bewertungskriterien*, *Handlungsoptionen* und *Reflexion* bei der Aufgabenentwicklung unterschieden (siehe Abbildung 2.2). Analog zum Göttinger Modell der Bewertungskompetenz wird dabei davon ausgegangen, dass die Anzahl verwendeter Kriterien in einer Entscheidungssituation einen Rückschluss auf die Komplexität bzw. die Elaboriertheit des Bewertungs-/Entscheidungsprozesses erlaubt (siehe auch Kapitel 2.5.1). Aufgaben für die Erhebung bei ESNaS werden in der Regel so erstellt und variiert, dass sie jeweils einer spezifischen Kombination aus Komplexität, kognitivem Prozess und Kompetenzteilbereich zuzuordnen sind (Hostenbach 2011: S. 45).

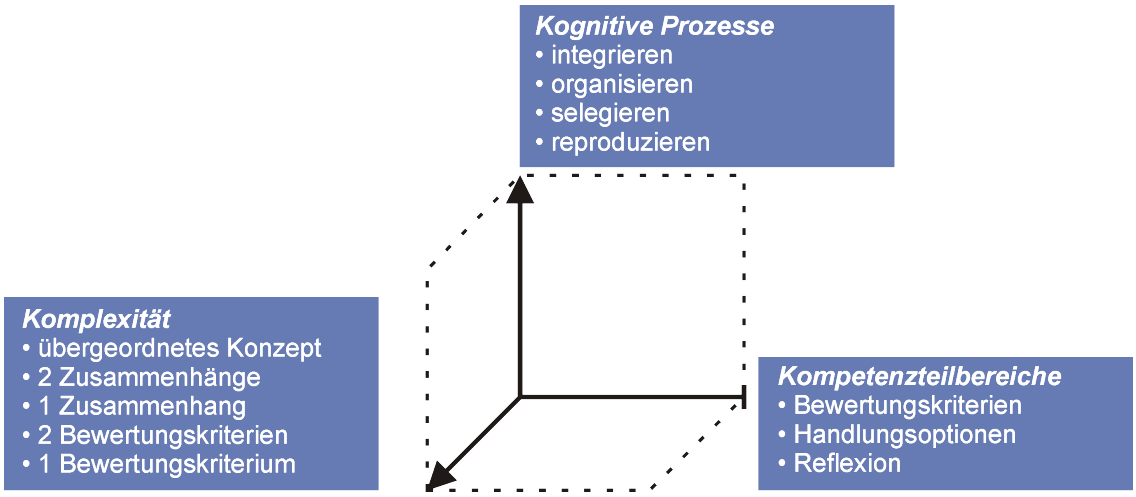

Abbildung 2.2: Dreidimensionales Kompetenzstrukturmodell für Bewertungskompetenz im Projekt ESNaS (nach Hostenbach 2011: S. 41).

Vergleich mit Göttinger Modell

Bestimmte Aspekte der Teilkompetenz *Bewerten, Entscheiden und Reflektieren* aus dem Göttinger Modell der Bewertungskompetenz sind auch im ESNaS-Modell für Bewertungskompetenz eingebunden. Die Teilkompetenz *Generieren und Reflektieren von Sachinformationen* wird hingegen nicht in das ESNaS-Modell mit aufgenommen, da sie dort bereits im Kompetenzbereich *Erkenntnisgewinnung* verortet ist, z. B. durch das Erkennen einer Problemstellung oder die Vergegenwärtigung der Wissensbasis (Hostenbach 2011: S. 45).

2.3 Socio-Scientific decision-making

Naturwissenschaft und Gesellschaft

Ein Pendant zu Bewertungskompetenz auf internationaler Ebene bildet das Konzept des *Socio-Scientific decision-making* (vgl. Ratcliffe 1997, Ratcliffe & Grace 2003). Geeignete Kontexte aus der Schnittstelle zwischen Naturwissenschaft und Gesellschaft werden dabei *Socio-Scientific Issues (SSI)* genannt. Sie werden beschrieben als „controversial social issues with conceptual and/or procedural links to science“ (Sadler 2011: S. 4). Aufgaben oder Probleme, die als SSI bezeichnet werden, sind nach Sadler (2004, 2011) stets durch folgende Aspekte gekennzeichnet:

Merkmale von SSI

- Umstrittene bzw. streitbare Sachverhalte mit Verbindungen zwischen Naturwissenschaft und Gesellschaft
- Offene Probleme ohne eindeutig vorgezeichnete Lösungen
- Beeinflusst von mehreren sozialen Faktoren, z. B. politischen, ökonomischen oder ethischen
- Entscheidungen sind nicht endgültig, die Datenlage kann sich ändern
- Quellen müssen hinterfragt werden.

Kontexte: global und/oder lokal

Sie sind somit gekennzeichnet durch offene Problemstellungen mit mehreren möglichen Handlungsoptionen, von denen keine „richtig“ oder „falsch“ ist. Diese Handlungsoptionen können dabei zwar durch naturwissenschaftliche Gesetzmäßigkeiten, Theorien oder Befunde beeinflusst sein, allerdings sind sie nicht allein durch naturwissenschaftliche Überlegungen vollständig bestimmt. Die Aufgabenkontexte und mögliche Handlungsoptionen sprechen zugleich soziale, politische, ökonomische und/oder ethische Aspekte an. Kontexte von SSI können entweder globaler Art (z. B. Klimawandel) oder lokaler Art sein (z. B. Standortwahl eines neues Kraftwerks) (Sadler 2011: S. 4). SSI sind in diesem Fall im Zusammenhang mit nachhaltiger Entwicklung zu bearbeiten (vgl. Robottom & Simonneaux 2012, Simonneaux & Simonneaux 2012, Tytler 2012).

weitere Konzepte

Mehrere Ansätze verfolgen das Ziel, naturwissenschaftlichen Unterricht in einen breiteren Kontext einzubetten, Schülerinnen und Schüler auf eine Beteiligung in damit verbundenen gesellschaftlichen Diskursen vorzubereiten und daraus Nutzen für die Gesellschaft insgesamt zu ziehen. Zu diesen Konzepten gehören vor allem *Science-Technology-Society (STS)* und *Science-Technology-Society-Environment (STSE)*. Das Konzept der SSI ist teilweise aus diesen hervorgegangen, es existieren daher gemeinsame Schnittpunkte aber auch unterschiedliche Schwerpunktsetzungen zwischen diesen Konzepten und SSI (vgl. Sadler 2011: S. 4).

SSI ab Klasse 6

Vorausgegangene Studien geben Hinweise darauf, dass SSI erst ab der 6. Klasse sinnvoll mit Schülerinnen und Schülern bearbeitet werden sollten, wenngleich auch Einzelstudien bekannt sind, bei denen diese bereits mit Grundschülerinnen und Grundschülern bearbeitet worden sind (Sadler 2011: S. 356). Im Rahmen der folgenden Studie werden daher Schülerinnen und Schüler ab Klasse 6 berücksichtigt (siehe Kapitel 4.4).

2.4 Herausforderungen durch Erneuerbare Energien als SSI und Schulrelevanz

Bildungsauftrag

Grundanliegen einer nachhaltigen Entwicklung im Kontext von Energieversorgung ist, die heute größtenteils eingesetzten fossilen Brennstoffe immer mehr durch Erneuerbare Energiequellen zu decken (vgl. Kapitel 1.3). Hieraus ergibt sich ebenfalls ein Bildungsauftrag für Schulen: Es ist überaus wichtig, junge Menschen mit dem nötigen Wissen über Aspekte der Energieversorgung auszustatten: Dies beinhaltet neben physikalischem Fachwissen auch ein Verständnis von Vorzügen und Risiken von fossilen und Erneuerbaren Energiequellen um als informierte und mündige Bürger reflektierte Entscheidungen treffen zu können (Gambro & Switzky 1999). Die Wichtigkeit wird auch dadurch untermauert, dass gerade der Energie-Kontext derjenige ist, der nahezu alle Facetten unserer Gesellschaft und unseres täglichen Lebens durchstreift: Die verlässliche Verfügbarkeit von Energie ist ein zentraler Baustein für die wirtschaftliche Entwicklung und die Sicherung von Lebensstandards (Bodzin 2012). Das Thema Energie vereint daher gleichermaßen auch interdisziplinäre Aspekte mit solchen, die primär physikalischen Ursprungs sind (Engström et al. 2011). Im Folgenden sollen nun ein paar internationale Arbeiten in Bezug auf deren ähnliche physikalische Kontextausrichtung und ihre Hauptergebnisse kurz vorgestellt werden:

kaum Verständnis von Energieumwandlung

In einer Studie hat Bodzin (2012) den Kenntnisstand von $N = 1043$ Schülerinnen und Schülern der 8. Klasse in Pennsylvania (USA) in Bezug auf die drei Subskalen *Energiebeschaffung, Energieerzeugung, -speicherung und -transport* und *Energienutzung und -einsparung* untersucht. Haupterkenntnisse dieser Untersuchung sind, dass das Verständnis

von Schülerinnen und Schülern zu Erneuerbaren Energiequellen zwar besser ist als das von fossilen, dass aber nur sehr wenige Schülerinnen und Schüler ein adäquates Verständnis von Energieumwandlungsprozessen (z. B. von Lage- in elektrische Energie bei einem Stausee) haben und wie viel Energie bei Tätigkeiten im Haushalt überhaupt benötigt wird (Bodzin 2012). Obwohl die Bildungspläne des Bundesstaats die Vermittlung dieser Kenntnisse vorsehen, werden sie von den in dieser Studie untersuchten Schülerinnen und Schülern klar verfehlt, was Bodzin (2012) unter anderem auch auf das Fehlen geeigneten Lernmaterials zurückführt.

Mängel in Lehrplänen

Die unzureichende Ausgestaltung bisheriger Lehrpläne wird auch international kritisiert: Chedid (2005) argumentiert, dass Bildung ein wesentlicher Faktor für technologischen Fortschritt ist, der in der amerikanischen Diskussion um den Mangel an technisch versierten Fachkräften erheblich unterschätzt wird. Er beschreibt Stand und Perspektiven, um die Bildung stärker in den Fokus der öffentlichen Diskussion in den USA zu rücken. Dass Wind- und Solarenergie als populäre Beispiele Erneuerbarer Energien im Unterricht an griechischen Schulen unterrepräsentiert sind, diskutieren Liarakou et al. (2009) und schlussfolgern, dass dieses Phänomen gar nicht primär auf das Wissen und die Einstellungen zu Erneuerbaren Energien von Lehrkräften zurückzuführen ist, sondern in Griechenland vielmehr durch Mängel in den Curricula für Schulunterricht und Ausbildung von Lehrkräften begründet liegt.

Bei der Untersuchung geeigneter Kontexte für das Unterrichten von Energie im Hinblick auf eine nachhaltige Entwicklung haben Engström et al. (2011) mehrere Themengruppen identifiziert, die nach einer Expertenbefragung hierfür sinnvoll erscheinen. Neben der Empfehlung eines möglichst großen Variation an Themen plädieren Sie auch dafür, die Kontexte durch Einstreuen von technischen Anwendungen und ökonomischen Betrachtungen langlebig in den Köpfen der Schülerinnen und Schüler zu verankern.

Verankerung in deutschen Lehrplänen

Das Bildungsziel einer nachhaltigen Entwicklung ist unter anderem in den Nationalen Bildungsstandards (siehe Tabelle 2.1) verankert. In Deutschland wird von Schülerinnen und Schülern am Ende des Physikunterrichts der 10. Klasse unter anderem gefordert, dass sie „alternative technische Lösungen auch unter Berücksichtigung physikalischer, ökonomischer, sozialer und ökologischer Aspekte [vergleichen]“ können (Standard B2 in KMK 2005c). Für Niedersachsen nimmt das Kerncurriculum für das Unterrichtsfach Physik eine differenziertere Aufstellung von prozess- und inhaltsbezogenen Kompetenzen für das Ende der Klassenstufen 6, 8, 10 und 12 vor, die neben vielen anderen Zielen auch Kenntnisse von Energieumwandlungsprozessen und Aspekte der Energieerhaltung, zur Einschätzung des häuslichen Energiebedarfs und zur Energieübertragung in Alltagssituationen vorgeben (NdsMK 2007, 2009). In Deutschland wurden adäquate Lernumgebungen für den Energie-

kontext zum Beispiel von Pahl et al. (2010) und Eilks et al. (2011) entwickelt und für den Einsatz im Unterricht bereitgestellt.

Projektbeispiel in den USA

Ebenfalls im Umfeld von SSI und Energieversorgung haben Rose und Barton (2012) einen Schüler und eine Schülerin ($N = 2$) in Michigan (USA) während einer nachmittäglichen AG beobachtet, wie sie sich das Thema um den Neubau eines Kraftwerks am Rande ihrer Heimatstadt erarbeiten, Informationen hierzu rezipieren und welche Aspekte und Hintergrundinformationen Einflüsse auf die Meinungsbildung der beiden Jugendlichen haben. Die beiden Autorinnen fordern, dass neben den Schülerinnen und Schülern natürlich auch die Lehrkräfte Erfahrung im Umgang mit Abwäge- und Entscheidungsprozessen vorweisen müssen, um diese an ihre Schülerinnen und Schüler weitergeben zu können. Sie heben zudem hervor, dass man erlernte wissenschaftliche Werkzeuge und Verfahrensweisen gerade auch dafür verwenden kann, Einzelinteressen hinter Werbebotschaften zu identifizieren und sachlich fundierte von politischen oder ökonomischen Interessen zu trennen.

Das Abwägen verschiedener Interessen und technischen Möglichkeiten ist damit ein zentraler Bestandteil eines umfassenden Bewertungs- und Entscheidungsprozesses.

2.5 Göttinger Modell der Bewertungskompetenz im Kontext nachhaltiger Entwicklung

Theoretische Herleitung

Das Göttinger Modell der Bewertungskompetenz beruht aktuell auf drei voneinander unabhängigen Säulen (siehe Abb. 2.3): *Kennen und Verstehen von Werten und Normen im Kontext nachhaltiger Entwicklung*, *Generieren und Reflektieren von Sachinformationen* und *Bewerten, Entscheiden und Reflektieren*. Die letzten beiden Teilkompetenzen wurden aus dem Rahmenmodell der Entscheidungsfindung von Betsch und Haberstroh (2005) (siehe Kapitel 1.1.1) abgeleitet (Eggert 2008, Eggert & Bögeholz 2006, 2010): Mit der *präselektionalen Phase* werden in diesem Modell Prozesse bezeichnet, die der Informationssuche und der Generierung von Optionen dienen, dies ist durch die Teilkompetenz *Generieren und Reflektieren von Sachinformationen* abgebildet (Gausmann et al. 2010). Die Teilkompetenz *Bewerten, Entscheiden und Reflektieren* ist im Modell von Betsch und Haberstroh (2005) durch die *selektionale Phase* repräsentiert, während der es zur Bewertung von Optionen und zu einer Entscheidung kommt. Diese Phase ist somit die zentrale Phase von Entscheidungsprozessen (Eggert 2008, Eggert & Bögeholz 2006, 2010). Erste Interventionsstudien zur Förderung von *Bewerten, Entscheiden und Reflektieren* werden von Gresch und Bögeholz (2013), Gresch et al. (2011) vorgestellt. Der Stand der Forschung wird in Bögeholz (2011) dargelegt.

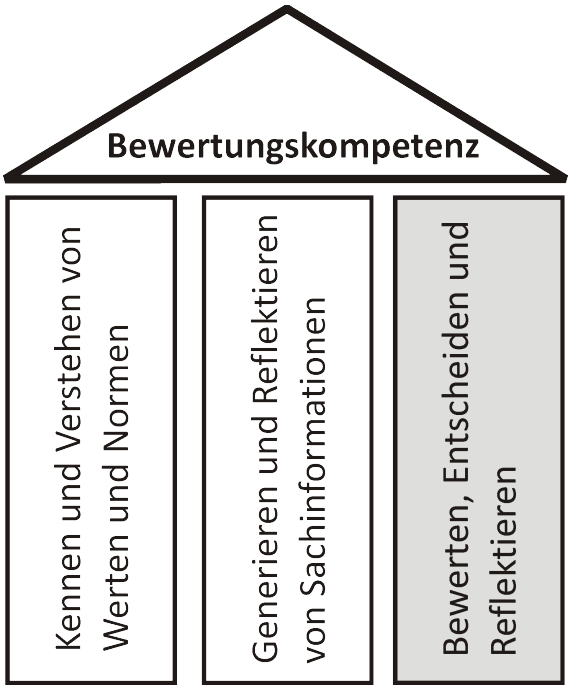

Abbildung 2.3: Göttinger Modell der Bewertungskompetenz im Kontext nachhaltiger Entwicklung mit den drei Teilkompetenzen als Säulen (nach Bögeholz 2011, aufbauend auf Eggert und Bögeholz 2006 und Bögeholz 2007).

2.5.1 Graduierungsüberlegungen bei Entscheidungsstrategien zum *Bewerten, Entscheiden und Reflektieren*

Eng verbunden mit den Kompetenzen von Schülerinnen und Schülern in Entscheidungssituationen ist die Wahl einer Entscheidungsstrategie. Auf den Überlegungen aus Kapitel 1.1.2 und 1.1.3 aufbauend haben Eggert und Bögeholz (2010: S. 234f) konkrete Graduierungsüberlegungen für die Teilkompetenz *Bewerten, Entscheiden und Reflektieren* im auch dieser Arbeit zugrundeliegenden Göttinger Modell der Bewertungskompetenz ausgearbeitet. Diese sind im Einklang mit den bereits dargestellten Ergebnissen früherer Forschungsarbeiten und erlauben die folgende konkretisierte Hierarchisierung (siehe Abbildung 2.4): Das unterste Level ist ein *spontanes* oder *intuitives Entscheidungsverhalten*, weil es eine spontane Wahl einer Option beschreibt, für die keine Vergleiche von Optionen angestellt wurden. Allenfalls wird stattdessen die gewählte Option lediglich unter Nennung von Kriterien im Nachhinein gerechtfertigt, was als *intuitiv-rechtfertigendes Entscheidungsverhalten* bezeichnet wird (Haidt 2001). Schülerinnen und Schüler, die z. B. Optionen zurückweisen, wenn ein Kriterium einen bestimmten Schwellenwert („cut-off“, Jungermann et al. 2005: S. 112) überschreitet, zeigen im Sinne dieser Hierarchisierung schon eine höhere Qualität von Entscheidungsfindungsprozessen, indem sie *non-kompensatorisch* vorgehen:

intuitiv

non-kompensatorisch

Sie vergleichen Kriterien bei verschiedenen Optionen, indem sie sukzessive vorgehen und nur einen Aspekt zur gleichen Zeit betrachten (Abelson & Levi 1985, Betsch & Haberstroh 2005, Jungermann et al. 2005). Das höchste Level von *Bewerten, Entscheiden und Reflektieren* für klassische Anforderungssituationen mit gleich legitimen Handlungsoptionen wird von *kompensatorischen* Strategien gebildet (Eggert & Bögeholz 2010), weil die Fülle von Informationen gleichzeitig betrachtet und gegeneinander abgewogen werden muss („trade-off“, Jungermann et al. 2005: S. 119) und dies nur bei ausgeprägteren kognitiven Fähigkeiten möglich ist (Abelson & Levi 1985, Betsch & Haberstroh 2005, Jungermann et al. 2005). Im Idealfall werden beim detaillierten Vergleichen von Optionen bereits einzelne Kriterien gewichtet. Zudem sind bestimmte Mischformen dieser beiden Vorgehensweisen denkbar und werden daher auch als Mischstrategien bezeichnet. Da Aspekte von beiden Strategien enthalten sind, rangieren diese im Rahmen der Logik der Graduierungsüberlegungen zwischen kompensatorischen und non-kompensatorischen Strategien. Denkbar wären z. B. der Beginn mit non-kompensatorischem Ausschluss einer Option, dann aber ein kompensatorischer Vergleich unter den verbleibenden Optionen. Diese Graduierungsüberlegungen sind im Einklang mit den in Kapitel 1.1.3 vorgestellten Analysen und der Anzahl verwendeter elementarer Informationsprozesse.

kompensatorisch

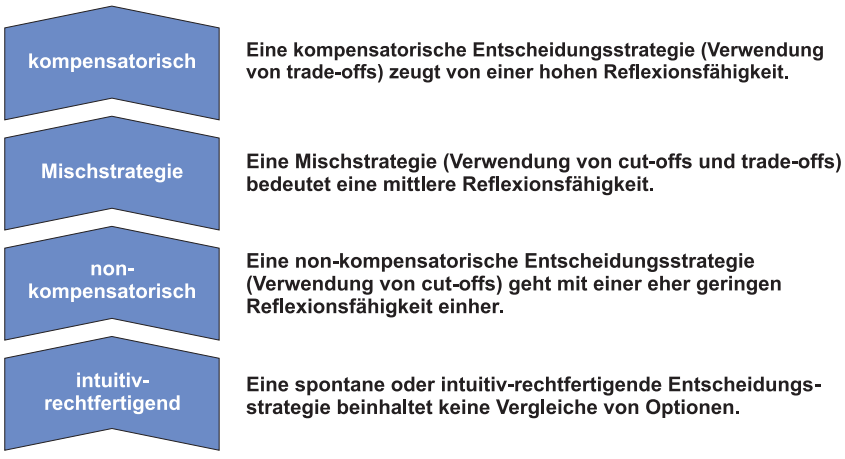

Abbildung 2.4: Graduierungsüberlegungen bei Entscheidungsstrategien (nach Eggert & Bögeholz 2010).

KAPITEL 3

Forschungsfragen

Die folgenden Forschungsfragen waren bei der Erstellung dieser Dissertation zielgebend:

3.1 Forschungsfrage I – Messbarkeit von *Bewerten, Entscheiden und Reflektieren*

Aufbauend auf den theoriegeleiteten Vorüberlegungen im vorangegangenen Kapitel werden Items für ein Testinstrument entwickelt, das geeignet sein soll, Bewertungskompetenz in der Teilkompetenz *Bewerten, Entscheiden und Reflektieren* bei Schülerinnen und Schülern zu erfassen und somit messbar zu machen. Die erste Forschungsfrage lautet daher:

Inwiefern ist es im Themenfeld Energiegewinnung, -speicherung und -nutzung möglich, mit dem neu entwickelten Testinstrument *Bewerten, Entscheiden und Reflektieren* im Kontext nachhaltiger Entwicklung zu messen?

Nach der Vorstellung des Testinstruments und seiner Items in Kapitel 4, der Beschreibung des Auswerteverfahrens in Kapitel 5 und weiterer Ergebnisse in Kapitel 6.1 werden Testgütekriterien in Kapitel 7.1 diskutiert.

3.2 Forschungsfrage II – Modellierung

Bei der Erstellung eines Testinstruments ist von vorneherein zu beachten, mit welchen Materialien und Methoden Ergebnisse erhoben und dargestellt werden sollen. Im Bereich der Kompetenzdiagnostik haben sich allgemein Verfahren der probabilistischen Testtheorie

bereits bewährt (siehe Kapitel 1 und 2). Auch im spezifischen Bereich der Bewertungskompetenz im Kontext nachhaltiger Entwicklung konnte das Rasch-Partial-Credit-Modell von Eggert (2008), Eggert und Bögeholz (2010) erfolgreich zur Modellierung der empirisch erhobenen Daten angewendet werden. Herausgefunden werden soll, inwiefern sich das eindimensionale Rasch-Partial-Credit-Modell für die Erklärung empirischer Daten zum *Bewerten, Entscheiden und Reflektieren* in physikrelevanten Kontexten einer Bildung für nachhaltige Entwicklung eignet. Die zweite Forschungsfrage lautet daher:

Sind die im Zusammenhang mit dem neu entwickelten Testinstrument empirisch erhobenen Daten im Rahmen eines eindimensionalen Rasch-Partial-Credit-Modells darstellbar?

Die Ergebnisse im Hinblick auf die erste Forschungsfrage werden in Bezug auf die Item-Fit-Indizes ausführlich in Kapitel 5.3 und zusammengefasst in Kapitel 6.2 dargestellt.

3.3 Forschungsfrage III – Umgang mit Kriterien bei Entscheidungen

Die im Testinstrument genannten Kriterien zu den jeweiligen Optionen liefern den Schülerinnen und Schülern eine Grundlage für ihren Entscheidungsfindungsprozess. Von Interesse ist daher, welche Kriterien sie bei ihren Bewertungen verwenden und wie sie diese in ihre Entscheidungen letztendlich einfließen lassen. Es ist zudem interessant zu erfahren, inwiefern sie Gewichtungen der Kriterien bei der Verwendung von Entscheidungsstrategien vornehmen (siehe Kapitel 1.1.3). Die dritte Forschungsfrage lautet deshalb:

Wie gehen Schülerinnen und Schüler mit Kriterien in Entscheidungssituationen um?

Ergebnisse zu dieser dritten Forschungsfrage werden in Kapitel 6.3 analysiert.

KAPITEL 4

Erhebungsmethodik

Die Testinstrumentenentwicklung zum *Bewerten, Entscheiden und Reflektieren* in physikrelevanten Bewertungskontexten einer Bildung für nachhaltige Entwicklung orientiert sind an der Operationalisierung von Eggert (2008), Eggert und Bögeholz (2006, 2010).

4.1 Offene vs. geschlossene Aufgaben

Kurzaufsatz

Bei Items mit offenem Antwortformat, wie z. B. der *Kurzaufsatzaufgabe*, bei der eine kurze schriftliche Abhandlung auf eine gestellte Frage erwartet wird, werden den Schülerinnen und Schülern eigenständige Leistungen abverlangt. Zum Zwecke der Leistungsmessung sind sie besonders geeignet, weil das Erzielen von zufällig richtigen Antworten durch Raten nicht möglich ist. Als Nachteil dieses Formats präsentiert sich dafür die anschließende Auswertung: Das Lesen und Nachvollziehen der Antworten von Schülerinnen und Schülern bedeutet einen relativ hohen Auswerteaufwand, den viele Forscherinnen und Forscher scheuen. Ein gut ausgearbeiteter Scoring Guide ist für die Codierer unerlässlich, um dem sonst möglicherweise auftretenden Problem einer geringeren Auswerteobjektivität zu begegnen (Pospeschill 2010: S.46). Large Scale Assessments, wie z. B. die PISA-Studie oder auch die ESNaS-Studie, setzen offene Antwortformate daher nur zurückhaltend ein und bedienen sich einfacher auszuwertender Erhebungsformate, wie z. B. Auswahlaufgaben.

Multiple Choice

Diese *Multiple-Choice-Aufgaben* (MC-Aufgaben) sind Aufgabenformate, bei denen mehrere Antwortmöglichkeiten bereits vorgegeben sind, und unter denen eine Schülerin bzw. ein Schüler die korrekte Antwortmöglichkeit auswählen soll. Um die Ratewahrscheinlichkeit zu reduzieren, sollten neben der richtigen Antwortmöglichkeit auch noch attraktive aber falsche Antwortmöglichkeiten, sogenannte *Distraktoren*, vorhanden sein. MC-Aufgaben sind zeiteffizient auswertbar und hinsichtlich ihrer Auswertung objektiv. Allerdings erfassen sie

oft nur Rekognitionsleistungen, Kreativität und herausgehobene Problemlösekompetenzen sind über MC-Aufgaben weniger gut erfassbar (Pospeschill 2010: S. 50).

4.2 Aufbau und Layout

Informations- und Antwortheft

Um den Schülerinnen und Schülern ein häufiges Hin- und Herblättern zwischen Aufgabenstellung und Antwortmöglichkeit zu ersparen, wurde ein zweiteiliges Testinstrument konstruiert: Das Informationsheft enthält die Aufgabenstellungen und Informationen zu den jeweiligen Aufgaben und das Antwortheft bietet Platz für die Antworten der Schülerinnen und Schüler. Beide Hefte sind im DIN A5-Format gedruckt, so dass sie sich voreinander auf den Tisch legen lassen. Querverweise untereinander sorgen dafür, dass die Schülerinnen und Schüler sich zurecht finden und die Aufgaben geordnet nacheinander bearbeiten.

zwei Entscheidungsaufgaben

Das Testinstrument beinhaltet drei Aufgaben. Die ersten beiden sind als *Entscheidungsaufgaben* angelegt: Schülerinnen und Schülern wird hier eine eigene Entscheidung für eine der jeweils dargebotenen vier Optionen abverlangt. Die Festlegung auf vier Optionen und vier bis fünf Kriterien hat sich bei vorausgegangenen Studien bewährt (Eggert 2008, Eggert & Bögeholz 2006, 2010), zumal man weiß, dass eliminative Strategien (wie EBA und LEX, siehe Tabelle 1.1) um so häufiger eingesetzt werden, je mehr Optionen vorgegeben werden (vgl. Jungermann et al. 2005: S. 136 und dortige Referenzen). Demgegenüber sollen die Schülerinnen und Schüler in der dritten Aufgabe, der *Reflexionsaufgabe*, keine eigene Entscheidungen treffen, sondern die vorgestellten Entscheidungswege von dritten Personen nachvollziehen und Verbesserungsvorschläge zu dem jeweiligen Vorgehen geben können. Ein struktureller Unterschied im Vergleich zu den von Eggert und Bögeholz (2006, 2010) eingesetzten Testheften ist in den Aufgaben zur Beschreibung der Strategien zu finden: Diese Teilaufgaben wurden nach der ersten Vorstudie (siehe Kapitel 4.6) in ein Multiple-Choice-Format abgewandelt.

eine Reflexionsaufgabe

Präsentationsformat

Bei allen Aufgaben wurde darauf geachtet, möglichst wenig Text zu verwenden und stattdessen auf eine komprimierte Präsentation der Aufgaben in Form von Tabellen zu setzen. Dies geschieht vor dem Hintergrund, dass der Einfluss von Lesekompetenz möglichst gering gehalten werden soll.

Kontextwahl

Bei der Konstruktion des Testinstruments und der Wahl des Kontexts sollte möglichst ein solcher gefunden werden, der von den Schülerinnen und Schülern auch ohne große physikalische Vorkenntnisse bearbeitet werden kann und in den man sich vergleichsweise schnell hineindenken kann. Jede Option sollte mindestens einmal im Vorteil gegenüber den anderen sein und es sollte sich zudem ein möglichst gleichmäßig verteiltes Bild unter den

Optionen bieten, damit es nicht zur voreiligen Wahl einer durch viele Vorteile hervorstechenden Option kommt. Kontexte wie „Kernenergie“ oder „Glühlampen“ sind ungeeignet, da sie zum einen durch die damalige öffentliche Debatte in den Medien sehr präsent waren und sich zudem die Gesetzgebung dahin entwickelt hat, dass beide Technologien nicht „zukunftsfähig“ sind (vgl. BGBl 2002, EC 2009). Schülerinnen und Schüler hätten damit eventuell eine vorgefertigte Meinung vertreten (siehe Kapitel 1.1.2) oder sich nicht auf die Informationen im Testheft eingelassen, was ebenfalls die Messung in Bezug auf einen höheren Anteil intuitiver oder non-kompensatorischer Strategien beeinflussen könnte. Unter Abwägung aller dieser Aspekte und in dem Bestreben einen für Schülerinnen und Schüler und zukünftige erwachsene Bürgerinnen und Bürger hochrelevanten und verständlichen Kontext vorzugeben, erfolgte die Wahl des Oberthemas *elektrische Energieversorgung* unter besonderer Berücksichtigung Erneuerbarer Energiequellen. Diese verbinden zudem das Bedürfnis nach Energie für das tägliche Leben mit dem Nachhaltigkeitsaspekt. Die Kontexte der einzelnen Aufgaben werden im Folgenden genauer dargestellt.

4.3 Kontext Energie

Energiekonzept

Das Energiekonzept ist eines der grundlegendsten Konzepte in den Naturwissenschaften, wenngleich es in den einzelnen Unterrichtsfächern mit unterschiedlichen Schwerpunktsetzungen behandelt wird. Thematisieren die Biologie und Chemie hierbei eher Stoffwechselreaktionen oder Energieumwandlungsprozesse, so befasst sich der Physikunterricht darüber hinaus noch mit der technischen Bedeutung von Energie für die Gesellschaft im Rahmen der Energieversorgung. In den letzten drei Jahrzehnten gab es viele Studien, die das Verständnis des Energiekonzepts im physikalischen Sinn (das weit über den Aspekt der elektrischen Stromversorgung hinaus geht und ebenso alle möglichen Energieformen und Umwandlungsprozesse betrachtet) bei Schülerinnen und Schülern untersucht haben (z. B. Duit 1986, Liu & McKeogh 2005, Neumann et al. 2013). Die herausgehobene Stellung des Energiebegriffs im Physikunterricht wird zum Beispiel auch im Niedersächsischen Kerncurriculum für das Fach Physik deutlich, indem er hier als „themenübergreifende Leitlinie“ bezeichnet wird (NdsMK 2007).

4.3.1 Windenergie als Vertreter Erneuerbarer Energien

Industrie-, Schwellen- und Entwicklungsländer sind gleichermaßen auf der Suche nach einer verlässlichen Energieversorgung. In Deutschland haben Weichenstellungen aus der Politik (wie u. a. der damalige „Atomausstieg“ (BGBl 2002), die „Laufzeitverlängerung“ (BGBl 2010) und deren Rücknahme (BGBl 2011)) einen verstärkten Bedarf nach Erneuerbaren

Energiequellen begründet. Grundtenor ist, Alternativen zu den derzeitig genutzten fossilen Energieträgern zu finden, da diese nur begrenzt zur Verfügung stehen und zudem schädliche Treibhausgase emittieren, die wiederum einen verstärkenden Effekt auf den Klimawandel haben (Quaschning 2011: S. 25ff). Windenergie hat sich zu einer gerade im norddeutschen Raum präsenten Erneuerbaren Energiequelle entwickelt und wird europaweit den Großteil des von der Europäischen Kommission vorgegebenen Ziels eines Anteils von 20% Erneuerbarer Energiequellen am Gesamtenergieverbrauch im Jahr 2020 beisteuern (insgesamt 17% allein durch Windenergie, EC 2007b). Deutschland war bis 2007 führend und ist 2011 hinter China und den USA immer noch das Land mit den größten Windenergieerträgen weltweit, zusammen erzeugen diese drei Länder 58% der global erzeugten Windenergie (GWEC 2012, WWEA 2011: S. 19). Windenergie

4.3.2 Kontext der ersten Entscheidungsaufgabe: Energieerzeugung durch Windenergieanlagen

Einzelne 6Windenergieanlagen werden meist zu Windparks zusammengeschaltet. Sie können dabei auf dem Festland („onshore“) oder auf dem Wasser („offshore“) installiert werden. Als Standorte werden den Schülerinnen und Schülern im Rahmen dieser Studie im Testinstrument vorgeschlagen: Nordsee (offshore), Ostsee (offshore), Küstenvorland und Mittelgebirge. Beim Auffinden eines geeigneten Standorts gibt es viele Faktoren, die berücksichtigt werden müssen. Einige werden den Schülerinnen und Schülern als Kriterien angeboten: Standortwahl

Das Kriterium „Windstärke“

Je nach Standort gibt es ein unterschiedlich ausgeprägtes Aufkommen an Wind: Bei den Offshore-Standorten ist dieser im Allgemeinen aufgrund der relativ glatten Meeresoberfläche meist kräftig und kontinuierlich. Auf dem Land ist der Wind im Allgemeinen weniger stark und unregelmäßiger, was unter anderem auf die rauere Oberfläche und den daraus resultierenden vermehrten Turbulenzen und Scherungen in der Atmosphäre zurückzuführen ist (Vaughn 2009: S. 37ff,59ff). Mit einem erhöhten Windaufkommen sind implizit natürlich eine höhere Effizienz und daraus resultierend höhere Gewinne für das investierende Unternehmen und ein höherer Windenergieertrag verbunden (Bilgili et al. 2011).

Das Kriterium „Kosten für Bau, Erreichbarkeit für Wartungsarbeiten“

Die Installation und der Betrieb von Offshore-Windenergieanlagen sind im Vergleich zu denen auf dem Festland höher. Zwar können die Bauteile recht einfach über Schiffe angeliefert werden, die Errichtung geeigneter Fundamente ist jedoch sehr aufwendig. Zudem werden bei der Offshore-Variante je nach Entfernung von der Küste zusätzliche Umspannwerke auf See benötigt. Auch die Erreichbarkeit für Wartungsarbeiten und Reparaturen ist im Vergleich zu den Festlandstandorten schlechter (Bilgili et al. 2011).

Das Kriterium „Sichtbarkeit und Lärmbelästigung“

Je mehr Windenergieanlagen errichtet werden, desto näher rücken sie auch an besiedelte Gebiete. Ob sich Anwohner durch die Anwesenheit einer Windenergieanlage gestört fühlen oder nicht, ist von einer Vielzahl von Faktoren abhängig. Lärmemissionen sind in den meisten Fällen zwar vorhanden, deren in *dB* gemessene Intensität steht aber nicht in Korrelation zum empfundenen Grad der Störung (Pedersen et al. 2007). Ein stärkerer Beitrag zum Empfinden einer Belastung scheint hingegen zu sein, dass Anwohner involviert werden wollen und es als störend empfinden, wenn der Bau einer Windenergieanlage in ihrer Nachbarschaft über ihre Köpfe hinweg entschieden wird. Eine verstärkte Bürgerbeteiligung verringert Widerstände und kann somit auch Planungs- und Bauzeiten erheblich reduzieren (Wolsink 2007)[1]. Den Einfluss weiterer Faktoren auf das Störungsempfinden untersuchten Bakker et al. (2012) unter anderem mit dem Ergebnis, dass sich Anwohner signifikant weniger gestört fühlen, wenn sie persönlich finanziell von einer Windenergieanlage profitieren. Offshore-Windenergieanlagen bringen diese Beeinträchtigungen in der Regel nicht mit sich, da sie fernab der Küste kaum zu sehen und nicht in der Nähe von besiedelten Gebieten errichtet werden können (Bilgili et al. 2011).

Das Kriterium „Naturschutz“

Auch Naturschützer kommen bei Windkraft, die sie als Alternative zu fossilen Energieträgern ja eigentlich begrüßen müssten, zu einem nicht einheitlichen Ergebnis, was den Einfluss von Windenergieanlagen auf Zugvögel betrifft: Unter Wissenschaftlern ist nach

1 Im Übrigen zeigt Wolsink (2007) auch positive Beispiele auf, wie nämlich durch mehrere Standortoptionen innerhalb einer Gemeinde die Bürgerbeteiligung und die Akzeptanz von Windenergieanlagen erhöht werden kann (im Gegensatz zu einer deutlich geringeren Akzeptanz, wenn die Entscheidungen „von oben“ kommen). Diskussionsgrundlage sind bei solchen Verfahren auch tabellarische Übersichten, die der Aufbereitung für Schülerinnen und Schüler in diesem Testheft sehr ähneln.

wie vor umstritten, ob Zugvögel durch Windenergieanlagen in ihrer Flugbahn gestört werden oder diese durch die kreisenden Rotorblätter gar getötet werden. Ein einheitliches Bild ist bisher vorliegenden Studien nicht zu entnehmen (Hötker et al. 2006: und Verweise darin).

Das Kriterium „Energieverlust beim Transport durch Stromleitungen“

Auch der Transport von elektrischer Energie von Windenergieanlagen zum Verbraucher ist ein Aspekt, den man bei den unterschiedlichen Standorten vergleichen kann. So lässt sich der in einer mit dem Netz synchronisierten Windenergieanlage erzeugte Windstrom direkt in das örtliche Niederspannungsnetz einspeisen und ohne größere Umwandlungsverluste von den Verbrauchern abnehmen. Im Gegensatz dazu muss Windenergie, die offshore erzeugt wurde, erst an Land zum Verbraucher transportiert werden. Da sich um die mit salzhaltigem Meerwasser umgebenen Leitungen aufgrund der Wechselspannung permanent magnetische Felder auf- und wieder abbauen, ist dieser Transportweg infolge der Blindleistung mit höheren Verlusten verbunden.

4.3.3 Kontext der zweiten Entscheidungsaufgabe: Energiespeicherung

intelligente Vernetzung

Ein großer Nachteil von Windenergie ist, dass sie nicht verlässlich produziert werden kann. Dies versucht man durch eine intelligente Vernetzung von weit entfernt liegenden Windparks und die Kombination mit Solarenergie zu umgehen: Regionen mit wenig Energieangebot könnten mit Strom aus Regionen versorgt werden, in denen gerade ein Überangebot erneuerbarer Energien verfügbar ist, was jedoch aufgrund nicht ausreichend ausgebauter Stromübertragungsnetze heutzutage nur bedingt möglich ist (VDE 2009, S. 167; dena 2012). Dennoch stellt sich die Frage, wie überschüssige Windenergie auf längeren Zeitskalen (z. B. Wochen) gespeichert und in windärmeren Zeiten wieder abgerufen werden kann. Die derzeitige Praxis sieht heutzutage (noch) so aus, dass Windenergieanlagen bei zu geringer Nachfrage abgeschaltet werden, da zu wenig Speicherkapazitäten zur Verfügung stehen und diese bisher kaum wirtschaftlich arbeiten (Leonhard et al. 2008: S. 45f).

Bedarf an Energiespeichern

Ebenfalls wichtig sind auch Ausgleichsmöglichkeiten im Tagesgang: Energie, die nachts im Überschuss vorhanden ist, wird gespeichert und in Zeiten höheren Bedarfs (z. B. während der Mittags- und Abendstunden) abgerufen. Verschiedene Energiespeichermedien werden daher in der zweiten Entscheidungsaufgabe im Testinstrument präsentiert, auch hier sind es nicht nur technische Kriterien, die bei einer Bewertung eine Rolle spielen können.

Speicher-technologien

Neben einer Speicherung in Form von Wasserstoff und der heutzutage gängigen Variante in Pumpspeicherkraftwerken kommen dabei auch die unterirdische Speicherung als Druckluft und die Speicherung in Akkumulatoren infrage. Alle Optionen sind technisch möglich und werden auch praktiziert. Physikalische und technische Informationen der jeweiligen Speichertechnologien wurden für die Konstruktion dieser Testaufgabe aus der Literatur zusammengestellt (Sauer 2010, UBA 2007, VDE 2009) und werden im Folgenden kurz beschrieben.

Das Kriterium „Wie gut kann Energie gespeichert werden?“

Die Wirkungsgrade der jeweiligen Energiespeichermöglichkeiten betragen ca. 40% bis 85% und wurden für dieses Testinstrument aggregiert mit den Begriffen „mittelmäßig“, „gut“ und „sehr gut“ angegeben (detailliert: Leonhard et al. 2008, Sauer 2010: jeweils S. 21f). Der Beschreibung in der Aufgabenstellung entsprechend wurden jeweils diejenigen Wirkungsgrade einer tagesbasierenden Speicherung gewählt. Für die Anwendung der geschilderten Speichertechnologien als Langzeitspeicher (mit Speicherzyklen in der Größenordnung von Wochen und/oder Monaten) können je nach Technologie auch geringere Wirkungsgrade erzielt werden.

Das Kriterium „Standortbedingungen“

Die Verfügbarkeit geeigneter Standortbedingungen stellt ein weiteres Kriterium an die verschiedenen Speichertechnologien dar: Einige Optionen sind an topographische Voraussetzungen gebunden, z. B. werden für die Druckluftspeicher unterirdische Salzkavernen oder stillgelegte Bergwerke und für ein Pumpspeicherkraftwerk geeignete Höhenunterschiede benötigt (VDE 2009: S. 158f). Für die Wasserstoffspeicherung können zur Erreichung großer Volumina oder Drücke ebenfalls Salzkavernen verwendet werden (VDE 2009: S. 54f), die Speicherung in oberirdischen Tanks ist jedoch praktisch überall möglich. Ebenso stellen Bleiakkumulatoren keine besonderen topographischen Anforderungen an einen Standort, sie lassen sich ebenfalls praktisch überall installieren (VDE 2009: S. 120).

Das Kriterium „Kosten für eine kWh“

Die entstehenden Kosten für die Speicherung betragen (stark vereinfacht) umgerechnet auf eine Kilowattstunde (kWh) ca. 6 bis 11 ct (Sauer 2010: S. 33f). Die Kosten beinhalten Investitions- und Betriebskosten für eine jeweils typische Speichergröße, zudem sind Schätzungen für kostendämpfende Weiterentwicklungen in der Zukunft bereits eingerechnet.

Ebenfalls Bestandteil der Kostenrechnung sind Szenarien einer Speicherung von Energie auf Tagesbasis.

Das Kriterium „Sicherheitsrisiken“

Etwaige Sicherheitsrisiken, wie z. B. die Gefahr eines Brands (beim Wasserstoffspeicher; DWV 2011), einer Überschwemmung (beim Pumpspeicherkraftwerk; Martin & Pohl 1996) oder die Verseuchung des Untergrunds mit Schwermetallen (beim Bleiakkumulator), werden ebenso für jede Option in stark reduzierter Form angegeben und bilden das letzte Kriterium einer jeden Option. Bei der Druckluftspeichermethode wäre ein unkontrolliertes Entweichen komprimierter Luft als Risiko denkbar, in der Aufgabenstellung werden jedoch (auch zum Erreichen einer möglichst gleichmäßigen Attributsverteilung zu dieser Option) keine Risiken angegeben, da selbst ein potientieller Störfall in seinem Ausmaß hinter Störfällen bei den anderen Technologien zurückbliebe. Die Dichtheit von Gaskavernen zur Einlagerung verschiedener Gase ist aktuell ein großes Thema und wird im Zusammenhang mit den geologischen Voraussetzungen für Kavernen diskutiert. Norddeutschland verfügt über vergleichsweise stabile Salzformationen, die für die Errichtung von Gaskavernen nahezu ideal und somit weniger risikoreich sind als in anderen Gegenden Europas (Gillhaus 2007: S. 11f).

4.3.4 Kontext bei der Reflexionsaufgabe: Energienutzung

Wie eingangs erwähnt, ist es nicht Ziel der Reflexionsaufgabe, die Schülerinnen und Schüler eine eigene Entscheidung treffen zu lassen. Vielmehr sollen sie die präsentierten Entscheidungswege Dritter nachvollziehen und Verbesserungsvorschläge dazu geben (Eggert & Bögeholz 2006, 2010). Ihnen werden Überlegungen einer Familie zum Kauf einer neuen Waschmaschine präsentiert. Die ersten beiden Kriterien, *Energie- und Wasserverbrauch* und *Restfeuchte nach Schleudergang*, spiegeln dabei die *Energieeffizienzklasse* und *Schleuderwirkungsgradklasse* wider, die auf dem Haushaltsgerät durch das EU-Energielabel beim Verkauf ausgewiesen sind (EC 1995: S. 23f). Als weitere Kriterien werden den Schülerinnen und Schülern *Transportweg und Herkunft*, *Anbieter* und *Preis* vorgegeben. Die Angaben zu den vorgestellten Verbraucherendgeräten basieren dabei jeweils auf konkret am Markt erhältlichen Waschmaschinen und unterstreichen damit den realitätsbezogenen Kontext der Aufgabe. Die von den Schülerinnen und Schülern zu identifizierenden drei Entscheidungsverhalten und -strategien werden jeweils verknüpft mit einer fiktiven Person dargeboten. Im Antwortheft befinden sich zu jeder der drei vorstellten Entscheidungswege eine MC-Frage (Items 11-13), mit deren Hilfe herausgefunden werden soll, ob Schülerinnen

Kaufentscheidung

und Schüler den präsentierten Entscheidungsprozess korrekt identifizieren und beschreiben können. Welche der präsentierten Entscheidungsprozesse am ehesten zu den ebenfalls präsentierten Wünschen der beiden fiktiven Personen im Testheft passt, sollen die Schülerinnen und Schüler in der folgenden Aufgabe darlegen (Item 14). In den weiteren Aufgaben (Items 15-17) sollen sie dann zusätzlich noch Verbesserungsvorschläge zu den jeweiligen Entscheidungsverhalten bzw. Entscheidungsstrategien unterbreiten (siehe Kapitel 1.1.2). Alle drei präsentierten Entscheidungsfindungsprozesse der fiktiven Personen im Testheft verfügen über weitere Optimierungsmöglichkeiten.

Das intuitiv-rechtfertigende Entscheidungsverhalten von „Marko“

Marko entscheidet sich spontan für eine der Waschmaschinen, ohne diese näher miteinander zu vergleichen. Er nennt im Anschluss drei von fünf möglichen Ausprägungen der Kriterien der gewählten Waschmaschine, wobei er Nachteile beschönigt darstellt und keine Gewichtung der Kriterien vornimmt.

Die non-kompensatorische Entscheidungsstrategie von „Bernd“

Bernd geht nach einer non-kompensatorischen Strategie vor (Elimination by Aspects (EBA), siehe Tabelle 1.1), indem er pro Kriterium immer die Waschmaschine eliminiert, die die schlechteste Ausprägung auf diesem Kriterium hat. Die Reihenfolge der Kriterien (die sonst auch maßgeblichen Einfluss auf die verbleibende und gewählte Option hat) wird dabei anhand zuvor festgelegter Wichtigkeiten betrachtet. Dabei bleibt nach drei Ausschlüssen eine Waschmaschine übrig, die dann allerdings auch nicht mehr hinsichtlich ihrer konkreten Ausprägungen betrachtet wird.

Die kompensatorische Entscheidungsstrategie von „Sabine“

Sabine entscheidet sich gemäß einer kompensatorischen Strategie (Weigthed Additive Rule (WADD), siehe Tabelle 1.1), nachdem sie die Charakteristika aller Waschmaschinen anhand einer Tabelle mit Punkten und diese entsprechend ihrer Wichtigkeit auch noch mit Gewichtungsfaktoren versehen hat (Attributmatrix, Jungermann et al. 2005: S. 43).

4.4 Prämissen der Aufgabenentwicklung

Authentizität

Um realitätsnahe Fragestellungen zu entwickeln, sollten die Aufgabenkontexte sowohl authentische Optionen beinhalten als sich auch ohne große physikalische Vorkenntnisse erschließen lassen, damit das Testinstrument für Schülerinnen und Schüler der Klassenstufen 6 bis 12 gleichermaßen geeignet sein kann. Die Attribute der Optionen wurden daher mit Fachliteratur recherchiert (siehe Kapitel 4.3.2 bis 4.3.4) und mit Experten rückgesprochen (D. U. Sauer, pers. Mitt., 11., 15. und 16. November 2009) und teilweise aggregiert dargestellt, um Schülerinnen und Schüler nicht zu überfordern.

reflektierte Entscheidungen

Von Schülerinnen und Schülern werden in diesem Testheft in der Regel *reflektierte Entscheidungen* erwartet. Dies bedeutet, dass nicht davon ausgegangen wird, dass sie bereits Routine in ähnlichen Entscheidungssituationen erworben haben, so dass sie unter Rückgriff auf die Erfahrung aus bereits vorgenommenen Entscheidungen *routinisiert* oder *stereotyp* vorgehen, dabei auf erlernte Verhaltensmuster zurückgreifen und demzufolge nur geringe kognitive Aufmerksamkeit einsetzen (Jungermann et al. 2005: S. 31-34,38). Es ist davon auszugehen, dass die Entscheidungssituationen für die Schülerinnen und Schüler neuartig sind; so neu, dass sie sich gewissenhaft den dargebotenen Informationen zuwenden, diese verarbeiten und in ihren Entscheidungsprozess einfließen lassen. *Konstruktive Entscheidungen* werden von den Schülerinnen und Schülern in diesem Testheft nicht erwartet, denn mögliche Optionen sind bereits vorgegeben und hinreichend dokumentiert, so dass sich Schülerinnen und Schüler sich diese nicht erst aus den gegebenen oder bei ihnen verfügbaren Informationen erschließen bzw. konstruieren müssen (Jungermann et al. 2005: S. 35f).

gleichlegitime Handlungsoptionen

Bei der Darstellung der Optionen bzw. deren Attribute wurde zudem darauf geachtet, dass keine Option dominant ist, also in allen ihren Ausprägungen besser als andere Optionen ist. Vergleichbar mit Eggert und Bögeholz (2010) werden den Schülerinnen und Schülern somit gleichlegitime Handlungsoptionen präsentiert. Dies geschieht vor dem Hintergrund, dass sonst eine Entscheidung zwischen den Optionen nicht schwerfällt und den Schülerinnen und Schülern kein echter Entscheidungsprozess abverlangt würde. Vielmehr wurde deshalb darauf geachtet, dass alle Optionen ausgeglichen präsentiert werden und man folglich durch Anwendung der Strategie *Equal Weight Rule (EQW)* (siehe Tabelle 1.1) zu keiner eindeutigen Entscheidung käme.

In einem kontinuierlichen und iterativen Prozess wurden die Aufgabenkontexte und deren Präsentation im Testinstrument in den Arbeitsgruppen der Göttinger Physik- und Biologiedidaktik mehrfach diskutiert und optimiert.

4.5 Validität der Testaufgaben

4.5.1 Curriculare Validität

Lehrplanbezug

Von curricularer Validität kann gesprochen werden, wenn sich Items eines Testinstruments in Übereinstimmung mit den in Lernplänen festgelegten Zielen befinden (Moosbrugger & Kelava 2008: S. 142). Ziel dieser Studie war es, ein Testinstrument mit Items zu entwickeln, welche an das in Kapitel 2.5 vorgestellte *Göttinger Modell der Bewertungskompetenz im Kontext nachhaltiger Entwicklung* anknüpfen und darüberhinaus auch noch eine hohe inhaltliche Passung zu den nationalen Bildungsstandards für Bewertungskompetenz in Physik aufweisen.

nationale Bildungsstandards

Wie bereits in Kapitel 2 expliziert, ist Bewertungskompetenz im Rahmen der Einführung nationaler Bildungsstandards auf Bundesebene definiert worden. Die für das Testinstrument entwickelten Items docken primär an den Standard B2 der nationalen Bildungsstandards für das Fach Physik an: „Schülerinnen und Schüler vergleichen und bewerten alternative technische Lösungen auch unter Berücksichtigung physikalischer, ökonomischer, sozialer und ökologischer Aspekte“ (KMK 2005c). Auf Bundesländerebene ergeben sich z. T. recht unterschiedliche Anknüpfungspunkte mit den Lehrplänen der einzelnen Bundesländer. Das in Niedersachsen derzeit gültige (und sich aktuell in der Revision befindliche) Kerncurriculum für die Naturwissenschaften der Sekundarstufe I spiegelt hinsichtlich des Unterrichtsfachs Physik nur vereinzelt Aspekte der Bewertungskompetenz im Sinne der nationalen Bildungsstandards wider (NdsMK 2007: S. 25ff). Der Kernlehrplan des Landes Nordrhein-Westfalen für die Sekundarstufe I greift hingegen Formulierungen aus den nationalen Bildungsstandards z. T. direkt auf und stellt explizit Gesellschaftsbezüge her. Für das Basiskonzept Energie formuliert er z. B., dass Schülerinnen und Schüler ausdrücklich „in die Lage [versetzt werden sollen], Bedeutung und Nutzen ebenso wie Gefahren der extensiven Energienutzung durch den Menschen einzuschätzen und verschiedene Möglichkeiten der Energiegewinnung, -aufbereitung und -nutzung unter naturwissenschaftlichen, wirtschaftlichen und ökologischen Aspekten zu vergleichen und zu bewerten sowie deren gesellschaftliche Relevanz und Akzeptanz zu diskutieren“ (NRWSW 2008: S. 22).

4.5.2 Inhaltsvalidität

Theoretische Fundierung

Theoretische Überlegungen, die Hinweise auf dieses Gütekriterium des Testinstruments geben, sind z. T. bereits bei der Vorstellung des Göttinger Modells für Bewertungskompetenz in Kapitel 2.5 vorgenommen worden. Der eingesetzte Mix aus offenen (Freitext-)

und geschlossenen (MC-Aufgaben) Aufgabenformaten ist auch unter Aspekten der Inhaltsvalidität bedeutend, da er auch auf die Erfassung und Abbildung unterschiedlicher Kompetenzniveaus abzielt (Moosbrugger & Kelava 2008: S. 140ff). Die Testaufgaben wurden ebenso bereits in ihrem Entwicklungsstadium in mehreren Runden mit Experten aus der Arbeitsgruppe der Göttinger Biologiedidaktik besprochen, die ihnen eine hinreichend gute Repräsentation von Bewertungsaspekten bescheinigen.

4.6 Vorstudien

zwei Vorstudien

Dem Prozess der Kontextfindung und Aufgabenkonstruktion schließen sich zwei Vorstudien an. Diese hatten zum Ziel, das Testinstrument zu erproben und für die Hauptstudie weiter zu verbessern. Dabei wurde zweistufig vorgegangen:

4.6.1 Studie Lauten Denkens

Lautes Denken

Auf qualitativer Ebene wurde das Testheft mit der Methode des Lauten Denkens (z. B. Deffner 1984, Ericsson & Simon 1984, Prüfer & Rexroth 2005) überprüft. Das Laute Denken ist eine etablierte Methode in der Problemlöse- und Denkpsychologie und hat auch Eingang in die Entscheidungspsychologie gefunden (Abelson & Levi 1985, Jungermann et al. 2005: S. 137). Die Methode ist dafür geeignet, kognitive Prozesse beim Entscheiden sichtbar zu machen und liefert wichtige Informationen zur Arbeitsweise von Schülerinnen und Schülern mit neu entwickelten Testheften (Betsch et al. 2011: S. 192f). Die wichtigsten Ergebnisse aus dieser Vorstudie (Meyer 2010, Sakschewski et al. 2013a) werden im Folgenden präsentiert:

Durchführung Studie Lauten Denkens

Überprüft wurde das Testheft im März 2010 bei $N = 14$ Schülerinnen und Schülern (10 weiblich und 4 männlich) aus den Klassenstufen 6 bis 12 eines niedersächsischen Gymnasiums. Für die Durchführung der Interviews und zur Sicherung einer hohen Durchführungsobjektivität wurde ein Manual mit Instruktionen entwickelt, das die Versuchsleiterin und der Versuchsleiter wortwörtlich vorgelesen haben. Die mitgeschnittenen Interview-Protokolle (von jeweils ca. 25 bis 45 Minuten Dauer) wurden transkribiert und mittels strukturierter Inhaltsanalyse (Mayring 2008) mit dem Textbearbeitungsprogramm MAXQDA analysiert. So wurden einzelne Aussagen der Schülerinnen und Schüler deduktiv (angelehnt an Eggert 2008) und induktiv einem entwickelten Kategorienbaum zugeordnet. Damit konnte erfasst werden, wie Schülerinnen und Schüler mit den einzelnen Optionen und Kriterien umgehen. Um den Prozess der Informationssuche nachvollziehen zu können, wurden die Schülerinnen und Schüler eingangs mit dem Verfahren des Lauten Denkens anhand von

zwei Beispielaufgaben vertraut gemacht. Sie wurden u. a. dazu aufgefordert, alle Fenster im Haus aufzuzählen und dies Raum für Raum zu beschreiben (Meyer 2010: S.XVI). Die Methode war allen Schülerinnen und Schüler schnell vertraut und sie konnten sie gut auf die im Testheft beschriebenen Kontexte anwenden (Meyer 2010: S.70).

Allgemeine Beobachtungen

Es konnte festgestellt werden, dass die erste Entscheidungsaufgabe im Vergleich zur zweiten für Schülerinnen und Schüler vermehrt Anknüpfungspunkte für zusätzliche, nicht im Testheft erwähnte, Aspekte bietet, was ein Hinweis darauf sein könnte, dass der Kontext der ersten Entscheidungsaufgabe (Energiegewinnung) den Schülerinnen und Schülern geläufiger ist als der zweite Kontext (Energiespeicherung). Fünf Schülerinnen und Schüler trafen auch bei der Bearbeitung der Reflexionsaufgabe zunächst eine eigene Entscheidung, anstatt sich in die vorgegebenen Entscheidungswege hineinzudenken und diese unvoreingenommen zu analysieren (Sakschewski et al. 2013a). Das Layout des Testinstruments wurde daher für die weiteren Erhebungen so verändert, dass eine deutlichere Trennung der Entscheidungsaufgaben im ersten Teil und Reflexionsaufgaben im zweiten Teil erkennbar wird.

Analyse der Entscheidungsstrategien

Bei der Analyse wurden auch alle erwarteten und in Kapitel 1.1.2 beschriebenen Entscheidungsstrategien (intuitiv-rechtfertigend, non-kompensatorisch und kompensatorisch) bei den untersuchten Schülerinnen und Schülern identifiziert: In den Entscheidungswegen wurde vornehmlich die Verwendung einer kompensatorischen bzw. einer Mischstrategie erkennbar. Non-kompensatorisch wurde in den Entscheidungsaufgaben zwei- bzw. viermal vorgegangen, drei bzw. zwei Schülerinnen und Schüler rechtfertigten lediglich ihre intuitiv getroffenen Entscheidungen. Dabei wurde beobachtet, dass die Schülerinnen und Schüler häufig in beiden Entscheidungsaufgaben dieselben Strategien anwenden. In acht Fällen wurde dieselbe Entscheidungsstrategie in beiden Entscheidungsaufgaben angewendet, die anderen sechs Personen wählten jeweils eine benachbarte Strategie, z. B. eine Mischstrategie in Aufgabe 1 und eine kompensatorische in Aufgabe 2. Letzteres kann als Argument für die Eignung der beiden Aufgaben zur Messung von *Bewerten und Entscheiden* gedeutet werden. Das recht häufige Vorkommen kompensatorischer und Mischstrategien steht in enger Verbindung mit dem Scoring: Eine Einordnung als kompensatorische Strategie kann z. B. schon erfolgen, wenn mindestens zwei Optionen im Hinblick auf mindestens zwei Kriterien miteinander verglichen werden und dabei Vor- und Nachteile der Optionen benannt werden.

Reflexionsaufgabe

Schwierigkeiten hatten die Schülerinnen und Schüler bei der (vollständigen) Beschreibung der Entscheidungswege Dritter. Das Erkennen und Beschreiben des intuitiv-rechtfertigenden Entscheidungsverhalten gelang nur drei Personen und scheint daher besonders schwierig zu sein (siehe Tabelle 4.1). Im Gegensatz dazu konnten 10 Schülerinnen und Schüler die

non-kompensatorische Strategie teilweise oder vollständig methodisch beschreiben und die Beschreibung der kompensatorischen Strategie gelang fast allen Schülerinnen und Schülern. Das Aufgabenformat im Testinstrument wurde daher für die Beschreibungen der dargebotenen Strategien von Freitextaufgaben auf MC-Fragen umgestellt.

Tabelle 4.1: Reflexion von Schülerinnen und Schülern über Entscheidungswege Dritter (Sakschewski et al. 2013a).

Beschreibung der Strategien Dritter durch Schülerinnen und Schüler			
Schüler/-in beschreibt den Entscheidungsprozess einer anderen Person...	intuitiv-rechtfertigend	non-kompensatorisch	kompensatorisch
...vollständig richtig	3	4	5
...teilweise richtig	0	6	8
...nur inhaltlich	11	3	1
...nicht	0	1	0

Zusammenfassung Studie Lauten Denkens

Die Vorstudie diente zudem dazu zu überprüfen, inwiefern das Informations- und Antwortheft missverständliche Wörter enthalten, die für Schülerinnen und Schüler der anvisierten Klassenstufen 6 bis 12 möglicherweise ein Hindernis für die Bearbeitung darstellen könnten. Dies kann mit den gewonnenen Daten ausgeschlossen werden. Das Testinstrument enthält für Schülerinnen und Schüler bewältigbare Herausforderungen. Auch der Umfang der berücksichtigten Kriterien bei den Entscheidungsaufgaben erscheint dabei weitestgehend unabhängig von der Klassenstufe der Schülerinnen und Schüler (siehe Tabelle 4.2).

4.6.2 Erste Fragebogenstudie

Durchführung Fragebogenstudie

In der Annahme, dass sich in den mittleren Altersstufen keine neuen Probleme oder Schwierigkeiten ergeben, wenn das Testinstrument an den beiden Randextremen getestet wird, wurden für die zweite Vorstudie $N = 37$ Schülerinnen und Schüler einer 6. ($n_6 = 23$) und 12. Klasse ($n_{12} = 14$) an demselbem niedersächsischem Gymnasium ausgewählt (keine Schülerinnen und Schüler haben an beiden Vorstudien teilgenommen). Auch in der zweiten Vorstudie kam ein Testmanual zur Sicherung einer hohen Durchführungsobjektivität zum Einsatz.

Erfahrungen für Haupterhebung sammeln

Die Vorstudie hatte zum Ziel zu schauen, ob Verbesserungsmaßnahmen aus der ersten Vorstudie (es wurden einzelne Wörter umformuliert, sowie zwei der Optionen in der ersten Aufgabe umbenannt, um sie eindeutig zuordnen zu können) die gewünschte Wirkung hatten und sich Verständnisprobleme bei den Arbeitsanweisungen im für die schriftliche Erhebung umgearbeiteten Testinstrument ergeben. Ebenso diente die zweite Vorstudie der organisatorischen Erprobung: Schülerinnen und Schüler haben nun ein komplettes Testpaket (inklusive dem *Lesegeschwindigkeits- und Verständnistest (LGVT)* schriftlich bearbeitet,

Tabelle 4.2: Nennung und Verwendung (positiv und negativ) von Kriterien beim Vergleich von Optionen in den Entscheidungsaufgaben in der Studie Lauten Denkens (nach Meyer 2010: S. 53,68).

Verwendung von Kriterien und Optionen in den Entscheidungsaufgaben									
		Aufgabe 1				Aufgabe 2			
Klasse, Schüler/-in[a]	Art[b]	Mittel-gebirge	Küsten-vorland	Nordsee (Offshore)	Ostsee (Offshore)	Wasser-stoff	Druck-luft	Pump-speicher	Blei-akku
6	pos.		W	W N L	L	S	K R	Q K	Q K
Maja	neg.	W	W	E B	N E B	Q K R	Q S	S R	S R
6	pos.								
Feldpony	neg.			N		R			R
6	pos.	E B	B	W N L	W L	S	R	K	Q S
Clarinette	neg.								
6	pos.		N L B	W N L	W	S	R K	Q K	Q S
Sternchen	neg.	W		B	E	R	Q	R	
8	pos.	E	N	W N	W	K	S	K	Q S K
Campino	neg.	N L	W L	E	N	Q S R	S	R	R
8	pos.		N E B				S K R		
Basketball	neg.		W L				Q		
8	pos.	E B	N B	W N L	W L	S	K R	K	Q S
Felix	neg.								
8	pos.	E	W N L	W N L	W N L	S	K R	Q K	Q S
Mirko	neg.	N L	E	B	N B	Q K R	Q S	S R	K R
10	pos.	E B	W N L B	W L	W L		R S	K	Q K
Rodario	neg.	W N L	E	B	N B	Q K R	Q S	S R	R
10	pos.			W N	W			K	Q
Neox	neg.			E B	N	R	Q	R	R
10	pos.			W N L				Q K	
Christina	neg.			E B		R		R	R
10	pos.	E B	N L E B	W N L	L		K		Q
Blatt	neg.	W N L	W	E B	N	K R	S	R	R
12	pos.	B		W N L			Q	K	
Mme Louise	neg.		L	B	N	Q S K R	S	R	R
12	pos.			W N L		S	R	K	Q S K
Lisa	neg.	N L	L	B	N	Q K	Q S	S	R

Dunkelgrau hinterlegte Zellen zeigen an, für welche Option sich die Schülerinnen und Schüler entschieden haben.
Die Abkürzungen der verwendeten Kriterien lauten:
Aufgabe 1: W=Windstärke, N=Naturschutz, L=Sichtbarkeit und Lärm, E=Energieverlust, B=Baukosten
Aufgabe 2: Q=Speicherqualität, S=Standortanforderungen, K=Kosten für eine kWh, R=Sicherheitsrisiken

[a] Selbstgewählter Alias der Schülerinnen und Schüler.

[b] Das jeweilige Kriterium wird positiv oder negativ zu einer Option genannt.

wie für die Haupterhebung geplant. Erfahrungen in Bezug auf z. B. die benötigte Gesamtzeit inklusive Testinstruktionen, Beschriftung der Umschläge oder Ein- und Austeilen der Umschläge waren von Interesse, um die Haupterhebung besser planen zu können und um einen reibungslosen Ablauf der Hauptstudie sicherstellen zu können.

Tabelle 4.3: Erprobung der MC-Fragen zum Beschreiben der Strategien: Anzahl der gewählten Attraktoren und Distraktoren von $N = 37$ Schülerinnen und Schülern.

Klasse	*Antwortmöglichkeit*	*Item 11* *int.-rechtfertigend*	*Item 12* *non-kompensatorisch*	*Item 13* *kompensatorisch*
6	1	2	3	3[a]
	2	12[a]	8	4
	3	0	6	9
	4	6	4[a]	4
	ungültig/keine	3	2	3
12	1	0	0	3[a]
	2	11[a]	4	4
	3	1	2	5
	4	2	8[a]	2
	ungültig/keine	0	0	0
Gesamtzahl richtiger Antworten		23 (68%)[b]	12 (34%)[b]	6 (18%)[b]

[a] korrekte Antwort
[b] aller gültigen Antworten

Erprobung neuer MC-Aufgaben

Erprobt wurde in dieser zweiten Vorstudie auch eine Neukonzeption der Aufgaben für die Beschreibung der Strategien in der Reflexionsaufgabe. Statt einer offenen Aufgabe kommt nun eine geschlossene MC-Aufgabe zum Einsatz. Das Ergebnis ist, dass die Distraktoren für die MC-Aufgabe größtenteils gut gewählt scheinen: In der Summe sind für die Klassenstufen 6 und 12 alle Felder in der Tabelle 4.3 durch mindestens eine Antwort besetzt. Allenfalls die hohe Anzahl richtiger Antworten beim Item 11 (Beschreibung des intuitiv-rechtfertigenden Vorgehens) ist ein wenig unglücklich.

Bezüglich des Erhebungsablaufs haben sich ebenfalls keine Probleme und Nachfragen der Schülerinnen und Schüler ergeben, so dass das Testinstrument für den Einsatz in der Hauptstudie geeignet scheint. Für die Durchführung der gesamten Erhebung sollten pro Testsitzung ca. 60 Minuten eingeplant werden.

4.7 Stichprobe

Zeitpunkt

Das entwickelte Testinstrument wurde im Rahmen einer Erhebung im November 2010 an drei verschiedenen Gymnasien in Niedersachsen in den Jahrgängen 6, 8, 10 und 12 eingesetzt. Von den insgesamt ausgeteilten 857 Testheften wurden 7 Testhefte von der

anschließenden Analyse ausgeschlossen, da sich 2 Schüler weigerten, den auf freiwilliger Basis durchgeführten Test mitzuschreiben, und bei 5 Schülerinnen und Schülern während der Erhebung beobachtet wurde, dass diese das Testheft nicht ernsthaft bearbeiten. Es wurden im Nachhinein keine Schülerinnen und Schüler aufgrund möglicherweise zu schlecht passender Personenparameter von der Studie ausgeschlossen.

Zusammensetzung

Für die Analyse wurden also die Antworten von $N = 850$ Schülerinnen und Schülern berücksichtigt, von denen $n_{weiblich} = 390$ Schülerinnen und $n_{männlich} = 435$ Schüler sind, bei $n_{unbekannt} = 25$ Personen ist die Geschlechtszugehörigkeit unbekannt (siehe Tabelle 4.4). Auch die Zusammensetzung in den einzelnen Jahrgangsstufen ist ungefähr ausgeglichen: In die Analyse wurden $n_6 = 219$ Testhefte aus der 6. Klasse, $n_8 = 229$ Testhefte aus der 8. Klasse, $n_{10} = 222$ Testhefte aus der 10. Klasse und $n_{12} = 180$[1] Testhefte aus der Oberstufe einbezogen. Die Wahl der Jahrgangsstufen folgte zum einen dem Wunsch, einen Querschnitt über das gesamte Spektrum der gymnasialen Schulbildung machen zu können und richtet sich zum anderen an der Unterteilung im Niedersächsischen Kerncurriculum (NdsMK 2007). Alle Schülerinnen und Schüler haben dasselbe Testheft bearbeitet. Schülerinnen und Schülern der Oberstufe wurde ein Testheft vorgelegt, in dem sie gesiezt wurden.

Tabelle 4.4: Stichprobenzusammensetzung getrennt nach Jahrgangsstufen und Geschlecht.

Zusammensetzung der Stichprobe					
Jahrgang	männlich	weiblich	keine Angabe	gesamt	ausgeschlossen[a]
6	105	106	8	219	5
8	128	94	7	229	1
10	109	108	5	222	0
Oberstufe[b]	93	82	5	180	1
Gesamt	435	390	25	850	7

[a] aufgrund Nichtteilnahme (freiwillig) oder nicht angemessenem Verhalten während der Erhebung

[b] hierunter 39 Schülerinnen und Schüler aus Jahrgang 11 und 141 Schülerinnen und Schüler aus Jahrgang 12

Ablauf

Zur Erreichung einer möglichst hohen Durchführungsobjektivität wurden alle Erhebungen von demselben Testleiter und Autor dieser Studie durchgeführt. Zusätzlich wurde ein Testmanual entwickelt und wortwörtlich vorgelesen, das den einheitlichen Ablauf jeder Erhebung und damit die Vergleichbarkeit untereinander garantiert. Einleitend wurde die Lesekompetenz der Schülerinnen und Schülern mittels des *Lesegeschwindigkeits- und -verständnistest für die Klassen 6 bis 12 (LGVT)* erhoben, um sie im weiteren Verlauf

1 Hierunter befinden sich auch 39 Schülerinnen und Schüler aus dem 11. Jahrgang. Die Schülerinnen und Schüler der gesamten Subgruppe „Oberstufe“ absolvieren bereits das zum Jahr 2011 in Niedersachsen neu eingeführte Abitur nach 8 Jahren („G8“).

von den Ergebnissen für *Bewerten, Entscheiden und Reflektieren* abgrenzen zu können. Die Erhebung fand jeweils während einer Doppelstunde statt, die Erhebungsdauer (inkl. Instruktionen) für LGVT und Testinstrument für *Bewerten, Entscheiden und Reflektieren* betrug etwa 60 Minuten.

Auswahl der Klassen

Die Auswahl der Klassen (jeweils ca. 3 pro Jahrgang und Schule) erfolgte quasizufällig durch die Schule. Dabei wurden je nach Schule z. B. diejenigen Klassen gewählt, in denen in der Erhebungswoche sowieso gerade Unterricht ausgefallen wäre. Minderjährige Schülerinnen und Schüler, die trotz mehrfacher Information im Vorfeld keine Elterneinverständniserklärung abgegeben hatten oder am Testtag nicht dabei hatten, durften an der Erhebung nicht teilnehmen. Sie erhielten kein Testheft und wurden von der jeweiligen Lehrkraft mit anderen Aufgaben beschäftigt.

Der Erhebungsphase gingen keine kontextrelevanten Ereignisse (Überflutungen, Reaktorunglücke, Eröffnung des Offshorewindparks „Alpha Ventus“[1]) voraus, die in den Medien sehr präsent waren und somit potentiell hätten Schülerinnen und Schüler in ihrer Entscheidung beeinflussen können.

1 Die Eröffnung des ersten deutschen Offshore-Windenergieparks erfolgte erst am 27.04.2010 (alpha ventus 2010).

KAPITEL 5

Auswertungsmethodik

5.1 Datenaufbereitung

Um die Antworten der Schülerinnen und Schüler statistisch analysieren zu können, wurden sie von mehreren Ratern zunächst codiert. Die Codierung erfolgte anhand des zuvor festgelegten Scoring Guides. Bei der anschließenden probabilistischen Datenanalyse (siehe Kapitel 5.3.2) konnten dann Zusammenlegungen von den ursprünglich vergebenen Kategorien gefunden werden, die es im Rahmen dieser Studie erlauben, die Daten in einem eindimensionalen Rasch-Partial-Credit-Modell abzubilden (Wu & Adams 2007: S. 57). Die ursprünglich vergebenen und die im Rahmen der Raschmodellierung angepassten Scores sind in den Tabellen 5.1 und 5.2 dargestellt.

5.1.1 Codierung der Entscheidungsaufgaben (Items 1-10)

Kriterien

Bei der Codierung der Antworten der Schülerinnen und Schüler zu den Entscheidungsaufgaben wurde von den Codierern erfasst, welche Kriterien die Schülerinnen und Schüler in ihre Bewertung haben einfließen lassen und ob diese positiv oder negativ mit den einzelnen Optionen in Verbindung gebracht werden. Der (ursprüngliche) Score 2 wird vergeben, wenn die Schülerinnen und Schüler für eine Option jeweils positive und negative Ausprägungen anführen. Werden nur positive oder nur negative Kriterienausprägungen zu einer Option genannt, entspricht dies dem (ursprünglichen) Score 1. Wird die Option gar nicht erwähnt, wird der Score 0 vergeben. Mit dem Ziel einer Modellierung im Rasch-Partial-Credit-Modell wurden nach inhaltlichen Überlegungen Kategorien zusammengelegt (siehe Tabelle 5.1). Das Vorgehen wird detailliert in Kapitel 5.3.4 beschrieben.

Tabelle 5.1: Scoring Guide für die Bewertung der Entscheidungsaufgaben.

Scoring Guide für Entscheidungsaufgaben				
			Score	
		Schüler/-in erwähnte...	ursprünglich	angepasst[a]
Option 1		keine Kriterien	0	0
Item 1	Item 6	nur pos. oder nur neg. Kriterien	1	0
Mittelgebirge	Wasserstoff	pos. und neg. Kriterien	2	1
Option 2		keine Kriterien	0	0
Item 2	Item 7	nur pos. oder nur neg. Kriterien	1	0
Küstenvorland	Druckluft	pos. und neg. Kriterien	2	1
Option 3		keine Kriterien	0	0
Item 3	Item 8	nur pos. oder nur neg. Kriterien	1	0
Nordsee (Offshore)	Pumpspeichersee	pos. und neg. Kriterien	2	1
Option 4		keine Kriterien	0	0
Item 4	Item 9	nur pos. oder nur neg. Kriterien	1	0
Ostsee (Offshore)	Bleiakku	pos. und neg. Kriterien	2	1
Gewichtung		kein gewichtetes Kriterium	0	0
Items 5 und 10		ein gewichtetes Kriterium	1	1
		mehrere gewichtete Kriterien	2	1

[a] nach ersten Raschmodellierungen und Prüfung der Item-Fit-Indizes (siehe Kapitel 5.3.4)

Vergleich von Optionen

Bei direktem Vergleich der Ausprägungen zweier Optionen („A ist in ... besser als B“) wird bei der hervorgehobenen Option A die Nennung eines positiven Attributs vermerkt, jedoch nicht automatisch die Nennung eines negativen Attributs bei Option B. Nur wenn eine negative Eigenschaft einer Option noch einmal explizit aufgeführt wird, wurde dieses Attribut als negativ mit der Option konnotiert gewertet.

Beispiele:

„Das Mittelgebirge, weil dort sehr wenig Vogelzug ist, der Energieverlust sehr gering ist und die Kosten nicht so hoch sind.“ *Ursprüngliche Scores 1, 0, 0 und 0 für die Items 1 bis 4. Beispiel für die Nennung nur positiver Attribute bei der Option Mittelgebirge. Testheft Nr. 21 (CLMI, Klasse 6).*

„[...] Beim Küstenvorland ist mittelstarker konstanter Wind, die Kosten für Bau etc. mittel und es sind nur wenig Ortschaften in der Nähe, allerdings gibt es starken Vogelzug und mittleren Energieverlust. [...]“ *Ursprüngliche Scores 0, 2, 0 und 0 für die Items 1 bis 4. Beispiel für das Aufführen positiver und negativer Aspekte bei der Option Küstenvorland. Testheft Nr. 058 (INDI16, Klasse 8).*

„Das Küstenvorland ist meiner Meinung nach am besten geeignet, weil dafür spricht, dass Wind stärker ist und gleichmäßiger als im Mittelgebirge. Dagegen

spricht die Sichtbarkeit und Lärmbelästigung und der Naturschutz. Diese beiden Kriterien können nur die Offshore-Windparks erfüllen, welche aber in Kosten und Wartung sowie im Energieverlust Defizite mit sich bringen. Somit bleibt als [Gesamtlösung] nur das Küstenvorland.“ *Ursprüngliche Scores 0, 2, 2 und 2 für die Items 1 bis 4. Beispiel für die Nennung von positiven und negativen Attributen bei den drei Optionen Küstenvorland, Nordsee (Offshore) und Ostsee (Offshore). Testheft Nr. 508 (CLMI, Klasse 10).*

Gewichtungen von Kriterien

Die Frage, ob Schülerinnen und Schüler Gewichtungen von Kriterien vornehmen, wurde ebenfalls von den Codierern erfasst und mit dem Score 1 versehen, wenn die Schülerinnen und Schüler ein Kriterium gewichten und mit dem Score 2 versehen, wenn sie mehrere Kriterien gewichten. Die Unterscheidung, ob ein oder mehrere Gewichtungen vorgenommen wurden, kann jedoch als Ergebnis der Raschmodellierung mit der Zusammenlegung der entsprechenden Kategorien ebenfalls als dichotomes Item behandelt werden (siehe Tabelle 5.1). Dies ist konsistent zu den Befunden von Eggert und Bögeholz (2010). Ein Herauf- oder Herunterstufen der Wichtigkeit von Kriterien wurde dabei gleichermaßen als Gewichtung gezählt:

Beispiele:
„[...] Tiere sind viel wichtiger als Strom. [...]“ *Ursprünglicher Score 1 bei Item 5. Beispiel für das Heraufstufen der Wichtigkeit des Kriteriums „Naturschutz“. Testheft Nr. 159 (SARE02, Klasse 6).*

„[...] Da ich persönlich jedoch Windräder weder bezüglich des Anblicks noch bezüglich des Geräusches (wobei ich mich da noch genauer informieren müsste) für störend halte, empfinde ich dieses Kriterium als am unwichtigsten. [...]“ *Ursprünglicher Score 1 bei Item 5. Beispiel für das Herunterstufen der Wichtigkeit eines des Kriteriums „Sichtbarkeit und Lärm“. Testheft Nr. 972 (KIMA25, Klasse 12).*

„[...] Der Vogelzug spielt für mich keine Rolle. Alles in Allem ist mir hoher gleichbleibender Stromgewinn wichtiger als gute Erreichbarkeit.“ *Ursprünglicher Score 2 bei Item 5. Beispiel für die Gewichtung mehrerer Kriterien. Testheft Nr. 509 (GUST, Klasse 10).*

„[...] Hierbei ist die Gewichtung der Kriterien von großer Bedeutung. An erster Stelle steht die Art des Windes [...]. Danach folgt [...] die Intensität des Vogelzugs [...]. [...]“ *Ursprünglicher Score 2 bei Item 5. Beispiel für die Gewichtung von mehreren Kriterien. Testheft Nr. 881 (HEFR, Klasse 12).*

In dieser Studie wird eine Gewichtung von Kriterien nur als solche berücksichtigt, wenn die Schülerinnen und Schüler in ihren Antworten erwähnen, dass ihnen manche Kriterien wichtiger/unwichtiger als andere sind (siehe Kapitel 5.1.1). Zusätzlich haben Schülerinnen und Schüler in wenigen Fällen (in Aufgabe 1 zweimal und in Aufgabe 2 dreimal, siehe kommendes Beispiel) speziell überlegt, ob sie einzelne Kriterien gewichten sollen oder nicht. Dies wurde ebenfalls als „Gewichtung“ gezählt, weil die Schülerinnen und Schüler hierzu explizit Überlegungen angestellt und dargelegt haben.

explizite Gleichgewichtungen

> Beispiel:
> *Schüler erklärt zunächst sein eigenes Punktesystem und schreibt dann:* „Allerdings lässt man bei diesem Verfahren außer Acht, dass die 5 Vergleichskriterien möglicherweise unterschiedlich gewichtet sein können, also alle Vergleichsaspekte nicht denselben Wert haben. [...]“ *Ursprünglicher Score 2 bei Item 5. Beispiel für eine explizite (Gleich-)Gewichtung von Kriterien. Testheft Nr. 914 (DOHA14, Klasse 12).*

5.1.2 Codierung der Reflexionsaufgaben (Items 11-17)

Die Reflexionsaufgabe besteht aus drei MC-Fragen und vier Freitextantworten. Während die Codierung der MC-Fragen und der ersten Freitextaufgabe (Items 11 bis 14) sich aufgrund des Aufgabenformats sehr einfach darstellt, waren bei den anderen Freitextantworten die von den Schülerinnen und Schülern gegebenen Verbesserungsvorschläge jeweils danach zu unterscheiden, ob sie überhaupt einen Verbesserungsvorschlag geben und dieser dann inhaltlicher oder strategischer Natur ist. *Inhaltlich* bedeutet in diesem Fall, dass der Verbesserungsvorschlag eng auf einzelne Kriterien bezogen ist, *strategische* Verbesserungsvorschläge haben eine von einzelnen Kriterien losgelöstere Sichtweise, die den Entscheidungsprozess in den Blick nimmt. Bei den strategischen Verbesserungsvorschlägen wurde beim Scoring zudem noch genauer danach unterschieden, ob sie nur in die richtige Richtung gehen („teilweise strategisch“) oder schon vollständig und umfassend sind. Die Ergebnisse der ersten Datenanalyse und die Prüfung vielfältiger Variationen legten eine Reduzierung der Kategorienanzahl von vier auf drei Kategorien nahe (siehe Kapitel 5.3.4). Im Folgenden wird daher die Codierung bei den einzelnen Verbesserungsvorschlägen noch einmal ausführlich dargestellt.

Beschreibung und Verbesserungsvorschläge zu Strategien

Item 15: Codierung des Verbesserungsvorschlags für die non-kompensatorische Strategie

In die Kategorie *teilweise strategisch* wurden Antworten von Schülerinnen und Schülern aufgenommen, die zwar eine strategische Ausrichtung haben, aber nicht vollständig zu

generelles Vorgehen

Tabelle 5.2: Scoring Guide für die Bewertung der Reflexionsaufgabe.

Scoring Guide für die Reflexionsaufgabe				
Item Nr.		Schüler/-in erwähnte...	Score	
			ursprünglich	angepasst[a]
11	MC; Beschreibung	falsche Antwort	0	0
	int.-rechtfertigend	richtige Antwort	1	1
12	MC; Beschreibung	falsche Antwort	0	0
	non-kompensatorisch	richtige Antwort	1	1
13	MC; Beschreibung	falsche Antwort	0	0
	kompensatorisch	richtige Antwort	1	1
14	Passendere	keine / Option 1-3	0	0
	Entscheidung	Sabine / Option 4	1	1
15	Verbesserungsvorschlag	keinen Vorschlag	0	0
	für	nur inhaltlichen Vorschlag	1	1
	non-kompensatorische	teilweise strategischen Vorschlag	2	1
	Strategie	voll strategischen Vorschlag	3	2
16	Verbesserungsvorschlag	keinen Vorschlag	0	0
	für	nur inhaltlichen Vorschlag	1	1
	kompensatorische	Vorschlag „nichts zu verbessern“	2	1
	Strategie	teilweise strategischen Vorschlag	3	1
		voll strategischen Vorschlag	4	2
17	Verbesserungsvorschlag	keinen Vorschlag	0	0
	für	nur inhaltlichen Vorschlag	1	0
	int.-rechtfertigendes	teilweise strategischen Vorschlag	2	1
	Verhalten	voll strategischen Vorschlag	3	2

[a] nach ersten Raschmodellierungen und Prüfung der Item-Fit-Indizes (siehe Kapitel 5.3.4)

Ende gedacht sind. Dies kann z. B. der Fall sein, wenn die Schülerinnen und Schüler zwar fordern, dass alle Kriterien berücksichtigt werden sollen, sie aber nur bis zu zwei Optionen diskutieren. Ein vollständiger Vergleich aller Optionen findet hier immer noch nicht statt. Erwähnen die Schülerinnen und Schüler in ihren Antworten, dass Bernd alle Kriterien und alle Optionen in seine Entscheidung mit einbeziehen soll, so wurde dies der Kategorie *vollständig strategisch* zugeordnet. Ebenso, wenn die Antwort der Schülerinnen und Schüler erkennen ließ, dass Bernd nach einem Ausschlussverfahren vorgegangen ist und erläutert wurde, weshalb dieses nicht alle Waschmaschinen mit einbezogen hat (siehe Tabelle 5.2).

Beispiele:
„Bernd hat vergessen, dass die Waschmaschine 3 einen weiten [Transport-] Weg hat.“ *Ursprünglicher Score 1 bei Item 15. Beispiel für einen inhaltlichen Verbesserungsvorschlag. Testheft Nr. 159 (SARE02, Klasse 12).*

„Er hätte den Rest der Kriterien auch mit einbeziehen müssen, als nur noch Waschmaschinen 3 und 4 übrig blieben.“ *Ursprünglicher Score 2 bei Item 15.*

Beispiel für einen teilweise strategischen Verbesserungsvorschlag. Testheft Nr. 822 (SIWA26, Klasse 12).

„Bernd hätte die WM nicht gleich ausscheiden lassen sollen, nur weil ihm ein Kriterium nicht passt. Er hätte alle Vor- und Nachteile aller WM vergleichen sollen." *Ursprünglicher Score 3 bei Item 15. Beispiel für einen vollständig strategischen Verbesserungsvorschlag. Testheft Nr. 229 (INMA19, Klasse 10).*

„Er geht nach dem Ausschlussprinzip vor. Am Ende bleibt einfach eine WM über. Die Waschmaschine die über bleibt, prüft er gar nicht nach seinen Vor- und Nachteilen." *Ursprünglicher Score 3 bei Item 15. Beispiel für einen vollständig strategischen Verbesserungsvorschlag. Testheft Nr. 869 (HEAD07, Klasse 12).*

Item 16: Codierung des Verbesserungsvorschlags für die kompensatorische Strategie

generelles Vorgehen

Der Kategorie *teilweise strategisch* wurden analog Antworten von Schülerinnen und Schülern zugeschrieben, die eine strategische Ausrichtung haben, aber nicht vollständig richtig zu Ende gedacht sind bzw. bei denen nicht durchgängig sinnvoll argumentiert wurde. Für die Kategorie *vollständig strategisch* gilt als Anforderung, dass die Schülerinnen und Schüler als Verbesserungsvorschlag geben, unwichtige Kriterien[1] mit dem Faktor 0 zu gewichten (also dieses Kriterium bei der Auswahl von Optionen nicht mit einzubeziehen) und/oder die Punktvergabe wird auf einem hohen Niveau generell kritisiert, weil sie subjektiv oder die Umsetzung der Information in Punkte nicht angemessen ist.

Spezialfall „alles richtig gemacht"

Anders als bei den anderen Strategien in den anderen Aufgabenteilen wurde hier zudem eine Kategorie eingeführt um Antworten von Schülerinnen und Schülern aufzunehmen, die beschreiben, dass Sabine alles richtig gemacht habe und es deshalb keinen Verbesserungsvorschlag geben sollte. Dies geschah vor dem Hintergrund, dass es ungerechtfertigt erschien, diese Einschätzung vom Niveau her unterhalb eines inhaltlichen Verbesserungsvorschlags anzusiedeln. Das im Testheft vorgestellte kompensatorische Vorgehen von Sabine ist bereits sehr komplex und ausgewogen, so dass Antworten von Schülerinnen und Schülern, die dazu raten nichts zu verändern weil bereits alles gut berücksichtigt sei, höher als ein inhaltlicher Vorschlag eingestuft sein sollte (siehe Tabelle 5.2).

Beispiele:
„Sabine könnte in ihrem Vorgehen etwas objektiver sein. Nicht nur die Energiemenge ist sehr wichtig, auch der Transportweg. Er sollte möglichst kurz sein."

1 Im Aufgabenheft erläutern drei fiktive Personen, welche Kriterien ihnen beim Kauf einer Waschmaschine wichtig und welche unwichtig sind.

Ursprünglicher Score 1 bei Item 16. Beispiel für einen inhaltlichen Verbesserungsvorschlag. Testheft Nr. 969 (CHJO30, Klasse 12).

„Ich finde, dass sie eine gute Möglichkeit gefunden hat die Waschmaschinen zu vergleichen. Vielleicht ist ihre Methode ein bisschen zeitaufwändig." *Ursprünglicher Score 2 bei Item 16. Beispiel für einen Vorschlag, dass sie „alles richtig gemacht" hat. Testheft Nr. 317 (GTJT14, Klasse 10).*

„Sie könnte eventuell die Wichtigkeit noch etwas genauer unterteilen, um so zu einem eindeutigeren Ergebnis zu gelangen." *Ursprünglicher Score 3 bei Item 16. Beispiel für einen teilweise strategischen Verbesserungsvorschlag. Testheft Nr. 964 (SUHO12, Klasse 12).*

„Sie könnte unwichtige Kriterien aus der Wertung weglassen und nur wichtige nicht so stark bewerten." *Ursprünglicher Score 4 bei Item 16. Beispiel für einen vollständig strategischen Verbesserungsvorschlag. Testheft Nr. 454 (SIJÖ16, Klasse 10).*

Item 17: Codierung des Verbesserungsvorschlags für das intuitiv-rechtfertigende Vorgehen

generelles Vorgehen

Als *teilweise strategisch* wurden in diesem Aufgabenteil Schülerinnen- und Schülerantworten gewertet, die entweder dazu raten alle Optionen zu betrachten oder eine Gewichtung von Kriterien vorzunehmen. Enthielt der Verbesserungsvorschlag beide Aspekte, wurde dieser der Kategorie *vollständig strategisch* zugeordnet.

Beispiele:
„Eine Waschmaschine aus Deutschland kaufen, um den CO_2-Ausstoß zu verringern." *Ursprünglicher Score 1 bei Item 17. Beispiel für einen inhaltlichen Verbesserungsvorschlag. Testheft Nr. 594 (RADI20, Klasse 8).*

„Er sollte bei jeder Waschmaschine alles beachten." *Ursprünglicher Score 2 bei Item 17. Beispiel für einen teilweise strategischen Verbesserungsvorschlag. Testheft Nr. 484 (BIRE18, Klasse 8).*

„Er betrachtet nur Waschmaschine 3, vergleicht nicht mit den anderen Maschinen und wägt Kriterien nicht nach ihrer Wichtigkeit ab." *Ursprünglicher Score 3 bei Item 17. Beispiel für einen vollständig strategischen Verbesserungsvorschlag. Testheft Nr. 917 (HIDI16, Klasse 12).*

5.1.3 Inter-Rater-Reliabilität

Gegencodierung

Die Antworten der Schülerinnen und Schüler wurden in zwei Phasen durch mehrere Rater codiert, die vorher instruiert und geschult wurden. Zur Qualitätssicherung wurden während des ersten Durchgangs ca. 40% der Testhefte von einem zweiten Rater gegencodiert. Aufgrund mäßiger Inter-Rater-Reliabilitäten bei den Items 5 und 10 (Gewichtung von Kriterien) und den Items 15 und 17 (Verbesserungsvorschlag für die non-kompensatorische Strategie und das intuitiv-rechtfertigende Verhalten) wurde der Scoring Guide mit den Erfahrungen aus der ersten Codierung noch einmal präzisiert. Mit den überarbeiteten Scoring Guides (siehe Tabellen 5.1 und 5.2) haben dann neue Rater die Testhefte im Hinblick auf die Items 5, 10, 15, 16 und 17 neu codiert, nachdem Sie wiederum im Umgang mit den revidierten Scoring Guides geschult wurden. In diesem zweiten Codierungsdurchgang wurden nun alle Testhefte durch jeweils einen zweiten Rater gegencodiert.

Fleiss' κ

Zur Berechnung eines Übereinstimmungsmaßes wurde *Fleiss'* κ (Fleiss 1971, Wirtz & Caspar 2002: S. 75ff) herangezogen. Im Gegensatz zu *Cohen's* κ ist es auch für mehr als zwei Rater zugelassen und berücksichtigt, dass ein Pool von Ratern vorliegt, aus dem zwar jede Aufgabe von gleich vielen aber nicht unbedingt von denselben Ratern codiert wird.

Tabelle 5.3: Inter-Rater-Reliabilitäten als Fleiss-κ-Werte.

Inter-Rater-Reliabilitäten als Fleiss-κ-Werte																
Itemnummer																
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Codierungsdurchgang 1 mit 40% Gegencodierungen																
,82	,76	,76	,79	,34	,81	,81	,74	,75	,65	,98	,99	1,0	,96	,33	,68	,26
Codierungsdurchgang 2 mit 100% Gegencodierungen																
				,54					,56					,61	,62	,57
Gemittelt über alle schwarz gedruckten Werte: Fleiss' $\overline{\kappa}_{Fleiss} = 0{,}77$																

Alle κ-Werte sind signifikant mit p<.001.

Über alle Items gemittelt konnte eine recht hohe Inter-Rater-Reliabilität von $\overline{\kappa}_{Fleiss} = 0{,}77$ erzielt werden (siehe Tabelle 5.3). Vor der Datenanalyse wurden die noch unterschiedlich codierten Antworten der Items 5, 10, 15, 16 und 17 einem Expertenurteil unterzogen, indem der Autor dieser Studie sich für die eine oder andere Kategorie entschieden hat.

Über eine Einordnung bzw. Interpretation des κ-Werts gibt es in der Literatur unterschiedliche Auffassungen. Fleiss (1981: S. 218) selbst bezeichnet κ-Werte im Bereich 0,40 bis 0,75 als annehmbar bis gut und höhere Werte als ausgezeichnet. Allgemein sollte κ möglichst Werte von $\geq$ 0,70-0,75 annehmen (vgl. u. a. Greve 1997: S. 111, Bortz und Döring 2002: S. 277, und ausführlich mit weiteren Verweisen Wirtz und Caspar 2002: S. 59).

5.2 Umgang mit Missing Data

5.2.1 Grundprobleme

Auswirkungen

Fehlende Daten (missing data) sind bei empirischen Erhebungen unweigerlich vorhanden, die Gründe für das Fehlen von Daten sind vielschichtig und Anlass genug, einen eigenen Forschungsstrang zu bilden. Zum einen reduzieren fehlende Daten die Stichprobengröße (speziell bei klassischen Verfahrensweisen wie dem listenweisen Ausschluss), woraus zugleich die Verlässlichkeit der geschätzten Parameter leiden kann. Zum anderen kann die Gesamtstichprobe, wenn systematische Unterschiede zum Fehlen von Daten führen, verzerrt werden.

Score „0“ oder nicht beantwortet?

Beim Umgang mit fehlenden Daten in Leistungstests sind zunächst einmal zwei Grundhaltungen denkbar und gegeneinander abzuwägen: Ein Vorgehen wäre, fehlende Antworten mit dem Score 0 zu versehen, wenn zugleich sichergestellt ist, dass die Schülerinnen und Schüler genug Zeit hatten, das Testheft zu beantworten. Schließlich haben Sie ja eine Frage nicht beantwortet. Die andere Vorgehensweise wäre, fehlende Antworten als „nicht beantwortet“ zu codieren und dies in den weiteren Analysen auf diese Weise zu berücksichtigen. In der Literatur gibt es hierzu unterschiedliche Philosophien. Der Grundtenor einiger Fallstudien, die die Auswirkungen von genau diesen unterschiedlichen Codierungsweisen auf den jeweiligen Datensatz untersuchen, verweisen darauf, dass die Unterschiede im Ergebnis zwar nur gering sind, das Codieren fehlender Antworten mit „0“ aber eher schadet, als dass es nützt. Fehlende Antworten sollten also lieber als „nicht beantwortet“ behandelt werden, da mit dem Score „0“ versehene Datensätze sich in Simulationen anders verhalten und größere Modellfehler verursachen (Hohensinn & Kubinger 2011, Shin 2009).

Klassifikation fehlender Daten

Einen generellen Überblick über den Umgang mit fehlenden Daten für Zwecke der psychologischen Forschung liefern Lüdtke et al. (2007). Für die Klassifikation fehlender Daten ist es zunächst relevant, ob diese „vollständig zufällig“ fehlen (MCAR – missing completely at random), „zufällig“ fehlen (MAR – missing at random) oder „nicht zufällig“ fehlen (MNAR – missing not at random), also sich weder den Kategorien MCAR und MAR zuordnen lassen. Bei Large Scale Assessments, wie z. B. dem in der PISA-Studie verwendeten Multi-Matrix-Design, das den Schülerinnen und Schülern unterschiedliche Testhefte aber mit teilweise überlappenden Aufgaben zufällig zuweist, werden fehlende Daten bereits aufgrund des Testdesigns als MCAR eingestuft und sind damit in der Regel unproblematisch (vgl. Baumert et al. 2001, Lüdtke et al. 2007: S. 105).

listenweiser Ausschluss

Sofern fehlende Daten nicht mit Sicherheit als MCAR klassifiziert werden können, sind klassische Verfahren wie der *listenweise Ausschluss* (wie z. B. von SPSS und WINMIRA

praktiziert) gefährlich, denn dies kann zu verzerrten Parameterschätzungen führen (vgl. Lüdtke et al. 2007: S. 107). Sind z. B. weniger fähige Schülerinnen und Schüler nicht in der Lage, Aufgaben in einem Testheft zu beantworten und weisen daher systematisch weniger Antworten auf als fähigere Schülerinnen und Schüler, so würden durch den listenweisen Ausschluss systematisch die weniger fähigen Schülerinnen und Schüler von der Analyse ausgeschlossen und die durchschnittliche Fähigkeit der Gesamtstichprobe überschätzt. Dieses Verfahren sollte nach Graham et al. (2003) nur bei einer Ausschlussquote von weniger als 5 Prozent angewendet werden. Ein anderes klassisches Verfahren stellt der *fallweise Ausschluss* (wie z. B. in ConQuest realisiert) von Daten dar. Hier werden alle jeweils verfügbaren Fälle für die Berechnung von Korrelationen, Item-Schwierigkeiten etc. verwendet. Es kommt so im allgemeinen zu einer geringeren Ausschlussquote als beim listenweisen Ausschluss, allerdings kann auch dieses Verfahren zu Verzerrungen führen, wenn die Daten nicht MCAR sind und die Stichprobengröße bei jeder Einzelauswertung eine andere ist (vgl. Lüdtke et al. 2007: S. 107). Andere klassische Verfahren wie *Gewichtung* scheiden aus, da hierzu nähere Informationen über die Schülerinnen und Schüler von Seiten der Schule (z. B. Anzahl der Mädchen und Jungen) nötig wären (vgl. Lüdtke et al. 2007: S. 108).

fallweiser Ausschluss

Neben diesen klassischen Verfahren und modellbasierten Verfahren, sollen hier noch *imputationsbasierte Verfahren* vorgestellt werden, da sie zur Abgrenzung der Ergebnisse dieser Studie gegen Einflüsse von fehlenden Daten verwendet werden. Die *einfache Imputation* beschreibt ein Verfahren, bei dem fehlende Werte z. B. durch den Mittelwert anderer Antworten in einem Test ersetzt werden. Der Mittelwert des Items bleibt beim Einsetzen von Werten dieser Methode also unverändert, allerdings ändert sich die Verteilung der Werte. Auch von diesem Verfahren raten Graham et al. (2003) daher ab, zudem sind in dieser Studie nur die ganzzahligen Codes 0, 1 und 2 verwendet worden, die auch nicht auf einer Intervallskala beruhen. Imputierte Werte wie z. B. 0,3 machen deshalb keinen Sinn und müssten auf 0 gerundet werden. Mit geringeren Problemen behaftet, allerdings auch nicht gänzlich unproblematisch, ist das Verfahren der *multiplen Imputation*, bei dem aufgrund der Antworten auf anderen Items mehrere „plausible Werte“ berechnet und mehrfach eingesetzt werden. Es werden somit mehrere vollständige Datensätze erzeugt, die dann entweder wieder zu einem Datensatz zusammengefasst (Lüdtke et al. 2007, Rubin 1987) oder getrennt voneinander mit Standardmethoden analysiert werden können. Von allen vorgestellten Verfahren, auch den weiteren in Lüdtke et al. (2007), scheint das Verfahren der multiplen Imputation auch unter Berücksichtigung aller Nachteile am geeignetsten. Lüdtke et al. (2007: S. 115f) sprechen von der „Überlegenheit der neueren Verfahren zur Behandlung von fehlenden Werten [in der methodischen Literatur]“, verweisen zugleich aber auch auf damit verbundene Kritiken („multiple Datenerfindung“, Rost 2005, S. 146).

imputationsbasierte Verfahren

Jedenfalls sei es „keine vertretbare Strategie, dieses Problem zu ignorieren“ und auch mit dem „weit verbreiteten Vorgehen des fallweisen Ausschlusses“ würde „das Problem der Behandlung fehlender Daten keinesfalls umgangen“ (Lüdtke et al. 2007: S. 116). Rubin (1987: S. 114) zeigt auf, dass bei unter 10% fehlenden Daten bereits wenige (ca. 5) Imputationen genügen, Graham (2012: S. 67) empfehlen bei bis zu 10% fehlenden Daten aufgrund der heutzutage gesteigerten Rechenleistung die Einbindung von 20 Imputationen in die jeweilige Berechnung.

Plausible Values

In manchen Analyseprogrammen (wie z. B. ConQuest) sind Techniken implementiert, die *plausible values* in Verbindung mit dem *expected a posteriori (EAP)-Schätzer* verwenden, besonders in Large Scale Assessments ist bei dieser Technik „eine elegante Methode des Umgangs mit missing data verbunden“ (Rost 2004: S. 316). Dabei werden aufgrund der zunächst empirisch ermittelten einzelnen Verteilungen der Items Personen„messwerte“ für jedes Item gezogen und anschließend zur Gesamtverteilung des zu messenden Konstrukts zusammengefasst. Die in den Statistikprogrammen angegebenen EAP-Personenfähigkeiten sind also das Ergebnis (ggf. mehrfach gezogener) statistischer Zufallswerte aus empirisch ermittelten Verteilungen (Rost 2004: S. 316), berücksichtigen dafür allerdings das Nichtvorhandensein vollständiger Daten.

5.2.2 Konkrete Umsetzung im Rahmen dieser Studie

Umfang

Im Sinne einer größtmöglichen Transparenz bei der Auswertung dieser Studie soll an dieser Stelle nun beschrieben werden, welche Überlegungen nun zum konkreten Umgang mit fehlenden Daten geführt haben. Von den 850 in die Auswertung einbezogenen Testhefte sind 753 (88,6%) vollständig ausgefüllt worden, bei 97 (11,4%) fehlen einzelne Antworten. Bezogen auf die Zahl der insgesamt erhobenen Daten ist festzustellen, dass für insgesamt 14.243 (98,6%) der 14.450 Datenfelder Antworten vorliegen, die Anzahl fehlender Daten beträgt also nur 207 (1,4%). Von diesen ist das Item 16 („Verbesserungsvorschlag für die kompensatorische Strategie“) am häufigsten von fehlenden Daten betroffen: 70 mal (8,2%) wurde es als „nicht bearbeitet“ codiert, ihm folgen mit 49 (5,8%) fehlenden Daten das Item 14 („Welche Entscheidung passt besser“), mit 44 (5,2%) das Item 17 („Verbesserungsvorschlag für das intuitiv-rechtfertigende Entscheidungsverhalten“) und mit 14 (1,6%) fehlenden Daten das Item 15 („Verbesserungsvorschlag für die non-kompensatorische Strategie“). Alle anderen Items haben Fehlquoten von deutlich unter 1% (siehe Tabelle 5.4). Aus den Erhebungsprotokollen ist zudem bekannt, dass in einer Erhebung in einem Kurs des 12. Jahrgangs sieben Schülerinnen und Schüler das Testheft nicht innerhalb der vorgesehenen 60 Minuten beenden konnten.

Auswirkung der Codierweise

Dass im Testheft fehlende Daten eher am Testende als am Testanfang auftauchen, ist hier allerdings nicht ein Effekt der Position der Items innerhalb des Tests (mit Ausnahme der eben erwähnten sieben Schülerinnen und Schüler), und auch nicht auf einen Mangel an Zeit bei der Bearbeitung der Items zurückzuführen, vielmehr muss es als ein Resultat des Codierungsprozesses angesehen werden: Bei den ersten beiden (Entscheidungs-)Aufgaben werden die ersten fünf Items zusammen codiert. Äußern sich die Schülerinnen und Schüler z. B. zu einer einzelnen Option nicht, wird dies gemäß den Codierungsregeln als Score „0“ gewertet, nicht als „nicht bearbeitet“. Die Items 1 bis 5 und 6 bis 10 wurden also nur dann als „nicht bearbeitet“ codiert, wenn der für die Beantwortung vorgesehene Platz im Antwortheft keine Eintragung enthielt. Die Items der dritten (Reflexions-)Aufgabe hingegen wurden einzeln vergeben. Äußern sich Schülerinnen und Schüler zu einer bestimmten Strategie nicht, kommt es hier also eher zu einer Einsortierung als „nicht bearbeitet“ als im ersten Teil des Testhefts. Generell wurden im gesamten Test Aufgaben nur dann als „nicht bearbeitet“ codiert, wenn keine Eintragungen vorhanden waren, also auch keine Striche/Durchstreichungen des vorgesehenen Antwortplatzes. Sobald z. B. durch einen einzelnen Strich oder irgendeinen Text erkennbar war, dass sich die Schülerin bzw. der Schüler der Aufgabe angenommen hat, wurde diese mit dem Score „0“ oder einem seiner Antwort entsprechenden höheren Score berücksichtigt.

Tabelle 5.4: Anzahl und Verteilung fehlender Daten. Insgesamt umfasst der Datensatz 14.450 Datenfelder ($N = 850$ Testhefte à 17 Items).

Anzahl und Verteilung fehlender Daten																	
	Itemnummer																
Klasse	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Gesamt[a]
Klasse 6	2	2	2	2	2	1	1	1	1	1	1	6	7	30	9	26	23	117
Klasse 8	0	0	0	0	0	0	0	0	0	0	0	0	0	12	3	13	11	39
Klasse 10	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	16	6	26
Oberstufe	0	0	0	0	0	0	0	0	0	0	1	0	0	4	1	15	4	25
Gesamt	2	2	2	2	2	1	1	1	1	1	2	6	7	49	14	70	44	207
in %	,2	,2	,2	,2	,2	,1	,1	,1	,1	,1	,2	,7	,8	5,8	1,6	8,2	5,2	1,4

[a] bezogen auf die Gesamtzahl von 14.450 Datenfeldern

multiple Imputation sinnvollste Methode

Voraussetzung für die Anwendung von multipler Imputation ist, dass die fehlenden Werte der Kategorie MAR zuzuordnen sind. Allerdings führen Lüdtke et al. (2007: S. 111) mehrere Simulationsstudien an die aufzeigen, dass das Verfahren der multiplen Imputation selbst bei fehlenden Werten der Kategorie MNAR – für die es also eigentlich nicht vorgesehen ist – anderen Vorgehensweisen wie z. B. dem fallweisen Ausschluss überlegen ist (vgl. Collins et al. 2001, Newman 2003, Schafer 1997, Sinharay et al. 2001).

In dieser Erhebung kann weder explizit ausgeschlossen werden, dass fehlende Daten MNAR sind noch mit Sicherheit davon ausgegangen werden, dass die fehlenden Daten MAR sind.

Es handelt sich allerdings wie oben bereits erwähnt um einen recht geringen Anteil an fehlenden Daten. Das Verfahren im Rahmen dieser Auswertung soll daher sein, dass Berechnungen in SPSS und WINMIRA zunächst mittels (die Stichprobe leider verkleinernden) listenweisen Ausschlusses vollzogen wurden. Berechnungen mit ConQuest wurden zunächst mit fallweisem Ausschluss durchgeführt. In einem zweiten Schritt wird dann jeweils gezeigt, dass die gewonnenen Ergebnisse robust gegen mittels multipler Imputation vervollständigte Datensätze sind. Da die fehlenden Datensätze insgesamt nur 1,2% des gesamten Datenmaterials ausmachen (siehe Tabelle 5.4), wurden in SPSS mittels des Analysemoduls „Multiple Imputation“ fünf verschiedene Imputationen realisiert. Mit den auf diese Weise fünf vervollständigten Datensätzen wurden dann eigene Raschmodellierungen durchgeführt und deren Ergebnisse dann zur Bestätigung der Ergebnisse herangezogen, die mit dem Original-Datensatz (ohne imputierte Daten) erlangt wurden.

5.3 Probabilistische Testtheorie

Anal4yse von Antwortmustern

Im Unterschied zur Klassischen Testtheorie werden in der probabilistischen Testtheorie Antwortmuster analysiert. Durch Kenntnis der Personenfähigkeit und z. B. Itemschwierigkeit kann innerhalb des gewählten Testmodells das Antwortverhalten einer Person bzw. die Wahrscheinlichkeit, mit der die Person ein Item richtig löst, vorausgesagt werden (Bühner 2012: S. 494). Gerade im Bereich der empirischen Bildungsforschung haben sich probabilistische Testmodelle etabliert, nicht zuletzt durch Large Scale Assessments wie den PISA-Studien (siehe Kapitel 2.1, Baumert et al. 2001: u. a.) im internationalen oder der ESNaS-Studie (siehe Kapitel 2.2.3, Hostenbach 2011) im nationalen Vergleich. Für *Bewerten, Entscheiden und Reflektieren* im Kontext nachhaltiger Entwicklung in der Biologie konnten Studien bereits zeigen, das die empirisch erhobenen Daten auf ein eindimensionales Rasch-Partial-Credit-Modell passen (Eggert 2008, Eggert & Bögeholz 2006, 2010).

5.3.1 Das Rasch-Modell

gemeinsame Skala

Aus der Rasch-Modellierung resultieren Itemschwierigkeiten und Personenfähigkeiten, die auf einer gemeinsamen Skala dargestellt werden können, der *Person-Item-Map*. Hier sind Items nach ihrer Schwierigkeit und Personen nach ihrer Fähigkeit angeordnet (siehe Abbildung 6.2).

Mathematische Beschreibungen bilden jedoch nicht exakt die Wirklichkeit und das reale Antwortverhalten von Personen ab. Man bedient sich daher Fit-Statistiken (siehe Kapitel

5.3.2) um zu überprüfen, inwieweit ein Modell zur Erklärung empirischer Daten angewendet werden kann und ggf. welche Anpassungen erforderlich sind, um die Modellgüte zu verbessern.

Das Rasch-Partial-Credit-Modell

auch Teilpunkte

Die im vorangehenden Kapitel kurz beschriebenen Grundzüge des Rasch-Modells basieren auf dichotomen Items und wurden von Masters (1982) um die Idee erweitert, in einer Aufgabe auch Teilpunkte (Partial Credits) für teilweise richtige Antworten vergeben zu können. Dies hat auch den Nebeneffekt, das in einem Test auch unterschiedliche Antwortformate zum Einsatz kommen bzw. in den einzelnen Aufgaben unterschiedlich viele Punkte vergeben werden können (Strobl 2010: S. 55). Die aufgabenspezifische Wahrscheinlichkeit, dass eine Person i beim Item j die Antwortmöglichkeit c wählt, beträgt

$$P(u_{ij} = c|\theta_i, \beta_j) = \frac{e^{c\theta_i - \beta_{jc}}}{\sum_{l=0}^{m_j} e^{l\theta_i - \beta_{jl}}}$$

wobei die β_j die während der Analyse berechneten Aufgabenschwierigkeiten und die θ_i die ebenfalls berechneten Personenfähigkeiten darstellen (Strobl 2010: S. 55). Trägt man diese Wahrscheinlichkeiten gegen die Personenfähigkeiten auf, erhält man die sogenannte *Item Characteristic Curve (ICC)* einer Aufgabe, wie sie z. B. in Abbildung 5.2 mit $m_{14} = 3$ Antwortkategorien dargestellt ist.

5.3.2 Itemselektion und Fit-Statistiken

Für eine Überprüfung des Testhefts sollen nun die Kriterien aufgestellt werden, nach denen die Items für die Modellierung in einem eindimensionalen Rasch-Modell ausgewählt wurden. Auf Grundlage dieser Angaben wurde die Revision des Scoring Guides (siehe Tabellen 5.1 und 5.2) vorgenommen. Im Folgenden finden sich jeweils zu jeder Größe die Erläuterung anerkannter Grenzwerte, die in dieser Studie erzielten Werte und die Beschreibung der Vorgehensweise.

Itemselektion

Beim Prozess der Itemselektion ist zunächst zu erwähnen, dass Erfahrungen aus vorangegangenen Studien (Eggert 2008, Eggert & Bögeholz 2010) bereits in die Erstellung des Testhefts eingeflossen sind (siehe Kapitel 4). Auf das Erstellen eines großen Itempools, aus dem in der Regel im Zuge der Itemselektion dann wieder ein Großteil der Items aufgrund nicht zutreffender Item-Fit-Indizes herausfällt, konnte somit verzichtet werden. Auch mit dem Ziel der Vergleichbarkeit wurde vielmehr die Struktur der Items in den

oben genannten vorausgegangenen Studien übernommen. Diese hatte sich in früheren Erhebungen bewährt, so dass die Items inhaltlich auf andere Kontexte in der Physik im Zusammenhang mit Energie umgearbeitet und auf diese Weise neue Items geschaffen wurden. Dieses Unterkapitel beschreibt daher diejenigen Anforderungen, an denen sich die neu entwickelten Items und der daraus resultierende Scoring Guide orientieren, um eine Modellierung im Rahmen eines Rasch-Partial-Credit-Modells zu ermöglichen.

Der Datensatz wurde mit den beiden Programmen WINMIRA und ConQuest analysiert. Beide Programme haben eine unterschiedliche Vorgehensweise und verwenden auch unterschiedliche Item-Fit-Indizes, von denen die wichtigsten und im Rahmen dieser Studie verwendeten im Folgenden beschrieben werden sollen.

iterativer Prozess

Der Prozess der Itemselektion ist ein iterativer Prozess: Nach dem Herauslassen einzelner Items oder dem Zusammenlegen von Kategorien sollte daher die Analyse komplett wiederholt werden, da sich die einzelnen Item-Fit-Indizes verändern (Wu & Adams 2007: S. 62,69). Es kommt zudem auf die Gesamtwürdigung der Items an: „Das Weglassen von Items mit Overfit könnte den Test seiner bestpassenden Items berauben – andere Items sind nicht so gut wie diese“ (Bond & Fox 2010: S. 241).

Q-Index (WINMIRA)

Die Q-Indizes stellen ähnliche Werte wie die Trennschärfen in der Klassischen Testtheorie dar und nehmen Werte zwischen 0 und 1 an. Sie bleiben gering, wenn Personen mit hohen Personenfähigkeiten auch höhere Item-Scores bekommen und umgekehrt (Rost & von Davier 1994). Zudem berechnet sich der Q-Index unabhängig von Korrelationen und Residuen[1] zwischen erwartetem und realen Abschneiden von Personen (Rost & von Davier 1994). Sie sind hilfreich um Items mit ungewöhnlichem Antwortmuster zu identifizieren: Würde die Antwort einer Person bei einem Item exakt ihrer Fähigkeit entsprechen, würde dieser Index den Wert 0 annehmen (Rost 1999), das Antwortmuster dieses Items entspricht dann dem eines Items mit maximaler Trennschärfe. Nimmt der Q-Index hingegen Werte nahe 1 an, sind die beobachteten Antwortmuster genau umgekehrt zu den nach dem Rasch-Modell erwarteten. Werte von ca. 0,5 kennzeichnen ein zufälliges Antwortmuster (Bühner 2006: S. 365f).

1 Residuen bezeichnen die Differenz zwischen der jeweils beobachteten Punktzahl einer Person auf einem Item und dem erwarteten Antwortverhalten einer Person mit derselben Fähigkeit auf diesem Item. Dabei erfolgt die Berechnung in 2 Phasen: Im ersten werden die Personenfähigkeiten und Itemschwierigkeiten zunächst aufgrund der empirischen Daten berechnet, bevor diese dann in der zweiten Phasen noch einmal mit den jeweiligen individuellen Antwortmustern jeder Person verglichen und daraus die Item-Fit-Indizes (wie z. B. der WMNSQ-Wert) berechnet werden (Bond & Fox 2010: S. 236f).

Allgemeine Verfahrensempfehlung: Vernünftige Q-Werte liegen in der Regel zwischen 0,1 und 0,3 (Bühner 2006: S. 366). Items mit Q-Werten über 0,35 wurden zur Aussortierung vorgemerkt, da sie nicht mehr gut zum Rasch-Modell passen.
Konkrete Handhabung in dieser Studie: Drei Items (siehe Tabelle 5.5) zeigen deutliche Auffälligkeiten bezüglich ihrer Q-Indizes, die beobachteten Q-Indizes der anderen Items lagen zwischen 0,17 und 0,32 (siehe Tabelle 6.7) und sind somit unauffällig.

Z_Q-Werte (WINMIRA)

Eine z-standardisierte (normalverteilte) Variation des Q-Index ist der Z_Q-Wert. Mit ihm kann man prüfen, ob ein Antwortmuster signifikant von dem unter dem Rasch-Modell erwarteten Antwortmuster abweicht. Der Z_Q-Wert nimmt positive Werte an, wenn das Item eher ungeeignet ist (Item-Underfit) und negative Werte an, wenn das Item eher „zu gut“ zum Modell passt (Item-Overfit). Der letztere Fall ist von beiden der unproblematischere, wenngleich ein zu guter Item-Fit auf Redundanzen im Test schließen lässt (Rost 2004: S. 373f).
Allgemeine Verfahrensempfehlung: Als zulässige Z_Q-Werte werden hier Werte im Intervall [-1,96; 1,96] betrachtet (Rost 2004: S. 374), interpretiert werden sollten aber nur signifikante Z_Q-Werte.
Konkrete Handhabung in dieser Studie: Ein Item-Overfit kann bei den Items 4, 6 und 7 (Umgang mit einzelnen Optionen in den Entscheidungsaufgaben) beobachtet werden, ein Item-Underfit beim Item 13 (Beschreibung der kompensatorischen Strategie, MC-Aufgabe) und den Items X1, X2 und X3 (siehe Tabelle 5.5). Alle anderen Z_Q-Indizes liegen im Normbereich und sind damit unauffällig.

Weighted Mean Square (WMNSQ, ConQuest) und dazugehöriger T-Wert

Der Weighted Mean Square (WMNSQ, auch „Infit“) stellt die zentrale Größe in der Item-Fit-Statistik dar. Er gibt darüber Aufschluss, inwieweit die empirisch beobachteten ICCs mit den nach dem Rasch-Modell geschätzten ICCs übereinstimmen. Der Wert wird berechnet, indem die einzelnen standardisierten Residuen zunächst quadriert und vor der Mittelwertberechnung noch mit ihrer Varianz gewichtet werden[1]. Er ist somit robuster

1 Liegen beispielsweise Itemschwierigkeit und Personenfähigkeit dicht beieinander, wird das Item stark streuen: Es gibt in etwa gleich viele Personen mit richtiger und mit falscher Antwort und deswegen eine hohe Varianz. Ist aber zum Beispiel die Itemschwierigkeit viel geringer als die Personenfähigkeit, so wird es fast nur Personen mit richtiger Antwort geben und die Varianz dementsprechend klein sein (und umgedreht).

gegen Personen mit extremen Personenfähigkeiten als der normale Mean Square (MNSQ, auch „Outfit“), da extreme Beobachtungen (die mit geringerer Varianz einhergehen) nicht mehr so stark ins Gewicht fallen (Bond & Fox 2010: S. 238f). Der WMNSQ-Wert kann ebenso standardisiert angegeben werden: Der ungefähr normalverteilte T-Wert macht dann eine analoge Aussage: Entsprechen die beobachteten Werte der Modellerwartung, nimmt der T-Wert einen Wert nahe 0 an. Werte außerhalb von [-1,96; 1,96] bzw. gerundet [-2;2] werden so interpretiert, dass sie eher nicht mehr die Modellerwartungen erfüllen (Bond & Fox 2010: S. 239).

Allgemeine Verfahrensempfehlung: Der WMNSQ sollte möglichst nahe bei 1 liegen, jedoch innerhalb des Intervalls [0,75; 1,33] (Bond & Fox 2010: S. 238f). Strengere Intervallgrenzen von [0,8; 1,2] werden von der OECD (2002: S. 105) vorgeschlagen. Der T-Wert sollte möglichst nahe bei 0 liegen, jedoch innerhalb von [-2; 2] (Bond & Fox 2010: S. 238f). Wu und Adams (2007: S. 66) empfehlen diejenigen Items für eine Löschung vorzumerken, deren WMNSQ „viel größer“ als 1 ist; WMNSQ-Werte unter 1 sollten zunächst beibehalten werden, bis andere Item-Fit-Indizes gegen eine Beibehaltung des Items sprechen. Wu und Adams (2007: S. 66) verweisen ebenfalls darauf, dass es aufgrund unterschiedlicher Stichprobengrößen schwierig ist, absolute Ober- und Untergrenzen für den WMNSQ anzugeben.

Konkrete Handhabung in dieser Studie: Alle Items hatten einen WMNSQ deutlich innerhalb des Intervalls [0,75; 1,33]. Bei den T-Werten nehmen die Items 3 und 4 (Umgang mit einzelnen Optionen in der ersten Entscheidungsaufgabe) die Werte -2,1 und -4,1 an (siehe Abbildung 5.1), was sich als Overfit interpretieren lässt.

Trennschärfe (ConQuest)

Unter *Discrimination* gibt ConQuest die klassische Trennschärfe aus, die als Produkt-Moment-Korrelation berechnet wird (Wu et al. 2007, Wu und Adams 2007: S. 64). Die Trennschärfe ist ein Maß dafür, wie gut ein Item zu der Skala passt, die aus den anderen Items gebildet wird.

Allgemeine Verfahrensempfehlung: Die Trennschärfe sollte mindestens $\geq 0{,}20$ sein, vorzugsweise $\geq 0{,}40$ (Wu & Adams 2007: S. 64). Die Empfehlung der PISA-Studie 2003 für die Trennschärfe lautet $\geq 0{,}25$ (OECD 2005b).

Konkrete Handhabung in dieser Studie: Drei Items (siehe Tabelle 5.5) weisen ein unzureichende Trennschärfe auf, alle anderen Items haben Trennschärfen von $\geq 0{,}25$.

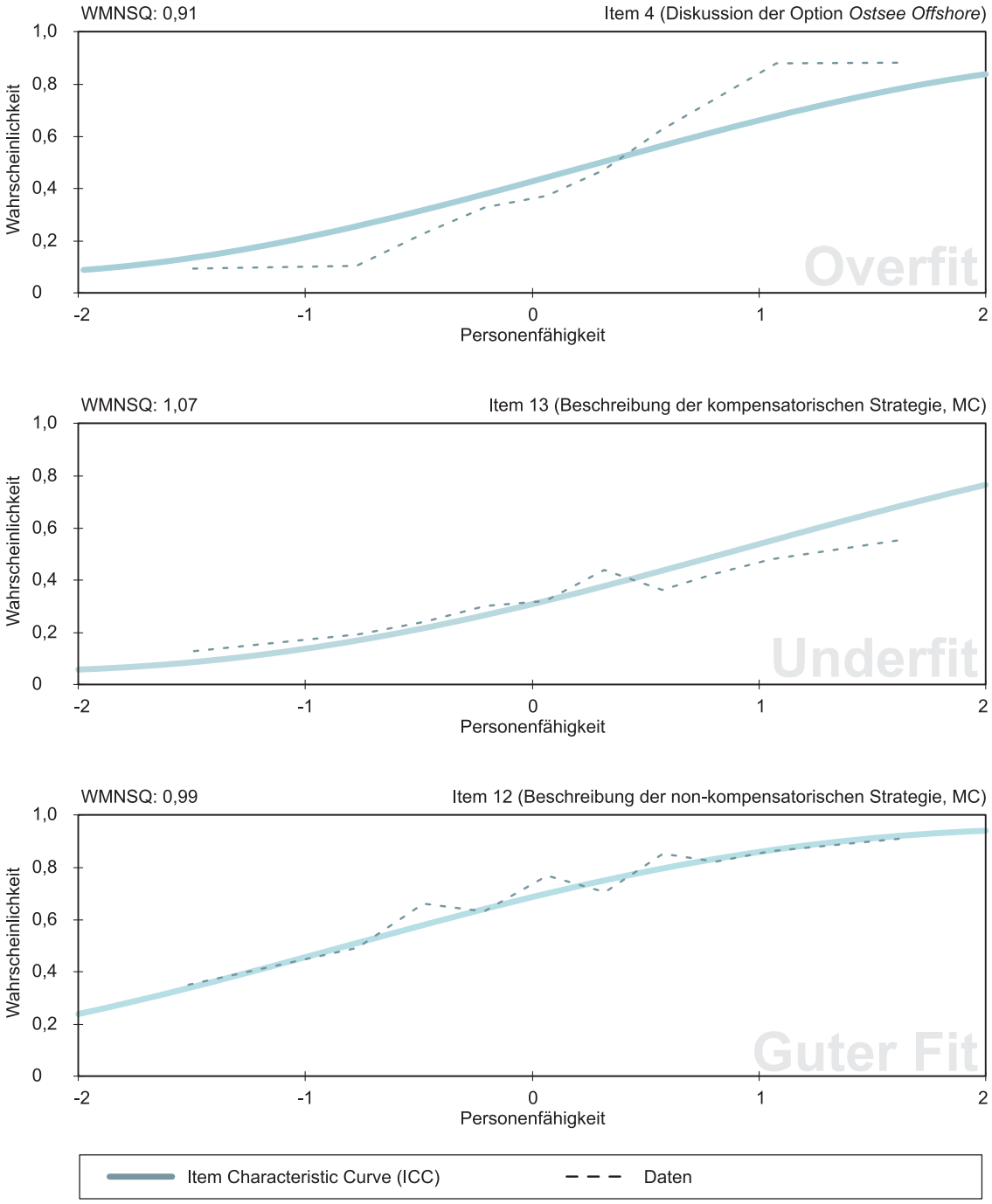

Abbildung 5.1: Item Characteristic Curves (ICC). Oben: Overfit (WMNSQ=0,91; T-Wert=-4,1) bei Item 4, „Daten passen zu gut zum Modell“. Mitte: Underfit (WMNSQ=1,07; T-Wert=2,0) bei Item 13, „Daten passen nicht so gut zum Modell“. Unten: Guter Fit (WMNSQ=0,99; T-Wert=-0,4) bei Item 12: „Daten passen gut zum Modell“.

Punkt-Biseriale Korrelation (Pt Bis, ConQuest)

Die Punkt-Biseriale Korrelation (Pt Bis) ist ein Kennwert dafür, wie die Antwort in einem dichotomen Item mit dem Summenscore (der aus der Addition aller Einzelitems gebildet wird) korreliert (Bühner 2006: S. 399f). Bei polytomen Items wird dazu eine dichotome Indikatorvariable berechnet nach dem Muster „hat die jeweils betrachtete Rubrik gewählt“ oder „hat eine andere als die jeweils betrachtete Rubrik gewählt“ (Nullrubrik), deren Punkt-Biseriale Korrelation mit dem Gesamtscore angegeben wird und damit so etwas wie die Trennschärfe auf Rubrikenebene darstellt (vgl. Lachmayer 2008: S. 85, Wu et al. 2007: S. 26,52). Daher sollte Pt Bis von den niedriger zu den höher bewerteten Rubriken ebenfalls monoton ansteigen und zudem bei den Nullrubriken nicht positiv sein, was nach Linacre (1998) ein Hinweis auf das Vorhandensein mehrerer Dimensionen ist.
Allgemeine Verfahrensempfehlung: Die Pt Bis sollte bei falschen Antworten bzw. Nullrubriken negativ sein. Die Pt Bis sollte bei polytomen Items geordnet sein und zu höheren Scores ansteigen (OECD 2005b: S. 123).
Konkrete Handhabung in dieser Studie: Fast alle Items weisen eine Pt Bis von 0,25 oder mehr auf, bis auf die Items 15 und 16 (vgl. Tabelle 6.7). Die Pt Bis bei Nullrubriken ist durchgängig negativ. Beim polytomen Item 15 sind die Pt Bis geordnet, bei den Items 16 und 17 fällt auf, dass die Pt Bis für den Score 2 jeweils wieder hinter die beim Score 1 zurückfällt und somit nicht mehr geordnet ist. Allerdings hängt sie immer noch positiv mit der Gesamtskala zusammen.

Thresholds (WINMIRA)

Thresholds werden auch als Schwellenparameter bezeichnet, weil sie diejenige Itemschwierigkeit angeben, ab der eine Person eine höhere Wahrscheinlichkeit hat, den (nächst)höheren Score zu erreichen. Ein Threshold bei einem dichotomen Item von +1 bedeutet also, dass Personen ab einer Personenfähigkeit von +1 mit höherer Wahrscheinlichkeit den Score 1 bekommen als den Score 0. Bei dichotomen Items sind die Thresholds daher identisch mit den Itemschwierigkeiten (*Item location* bei WINMIRA), bei polytomen Items ergibt erst der Mittelwert aller Thresholds die Itemschwierigkeit (Bühner 2006: S. 364f).
Allgemeine Verfahrensempfehlung: Die Thresholds sollten im Intervall [-3;3] angesiedelt sein (Baker 2001: S. 6). Zudem sollten die Schwellenparameter geordnet sein: Sukzessiv ansteigende Thresholds zu höheren Scores.
Konkrete Handhabung in dieser Studie: Dies war ein Hauptkriterium, nach dem bei der Zusammenfassung der Kategorien/Scores (siehe Kapitel 5.3.4) vorgegangen wurde. Immer kombiniert mit inhaltlichen Überlegungen wurden die Zusammenlegungen durch systema-

tische Variationen so gewählt (siehe Tabelle 5.7), dass die Thresholds allesamt im Intervall [-3; 3] liegen und geordnet sind, also zu höheren Scores ansteigen.

Item-Deltas (ConQuest)

Die Item-Deltas bei ConQuest beschreiben ein ähnliches Konstrukt wie die Thresholds bei WINMIRA, wenngleich die berechneten Grenzen in beiden Programmen (neben den unterschiedlichen Fallzahlen[1] (siehe Kapitel 5.2) auch noch aufgrund unterschiedlicher Algorithmen) nicht dieselben sind. Nach Wu und Adams (2007: S. 41ff) werden die Item-Deltas als Indikatoren für die Item-Schwierigkeit herangezogen und beschreiben (analog zu den Thresholds bei WINMIRA) diejenige Personenfähigkeit, ab der der jeweils nächsthöhere Score der wahrscheinlichere ist. Dies ist gleichzusetzen mit den Schnittpunkten der berechneten jeweiligen ICCs (siehe Abbildung 5.2 oben).
Allgemeine Verfahrensempfehlung: Die Item-Deltas sollten im Intervall [-3; 3] angesiedelt sein (Baker 2001: S. 6). Zudem sollten die Schwellenparameter geordnet sein: Sukzessiv ansteigende Deltas zu höheren Scores.
Konkrete Handhabung in dieser Studie: Dies war ein Hauptkriterium, nach dem bei der Zusammenfassung der Scoring-Kategorien (siehe Kapitel 5.3.4) vorgegangen wurde. Immer kombiniert mit inhaltlichen Überlegungen wurden die Zusammenlegungen durch systematische Variationen explorativ so gewählt (siehe Tabelle 5.7), dass die Item-Deltas möglichst im Intervall [-3; 3] liegen und geordnet sind, also zu höheren Scores ansteigen. Als etwas zu schwierig zeigt sich hier der Itemstep 17.2, der ein Item-Delta von 3,17 aufweist.

Item-Thresholds (ConQuest)

Die Item-Thresholds (gelegentlich auch als *Thurstonian Thresholds* oder *Gammas* (γ) bezeichnet) stellen eine andere Art von Schwellenwerten dar. Mit ihnen wird jeweils diejenige Personenfähigkeit beschrieben, ab der die Wahrscheinlichkeit für das Erreichen des jeweiligen Scores oder eines höheren Scores 50% beträgt (Wu & Adams 2007: S. 50). γ_1 bezeichnet also diejenige Personenfähigkeit, ab der mit einer Wahrscheinlichkeit von 50% mindestens der Score 1 erreicht wird (vgl. Abbildung 5.2 unten). Die Item-Thresholds können somit als Score-Schwierigkeiten bezeichnet werden (Wu & Adams 2007: S. 51).

1 Grundsätzlich arbeitet WINMIRA mit einem listenweisen Datenausschluss, ConQuest hingegen mit einem paarweisen Fallausschluss. Jedoch ergeben sich auch unterschiedliche Grenzen, wenn die Analyse nur mit den vollständigen Datensätzen ($n = 718$, siehe Tabelle 6.7) oder mittels multipler Imputation vervollständigten Datensätzen (siehe Tabelle 5.5) durchgeführt wird.

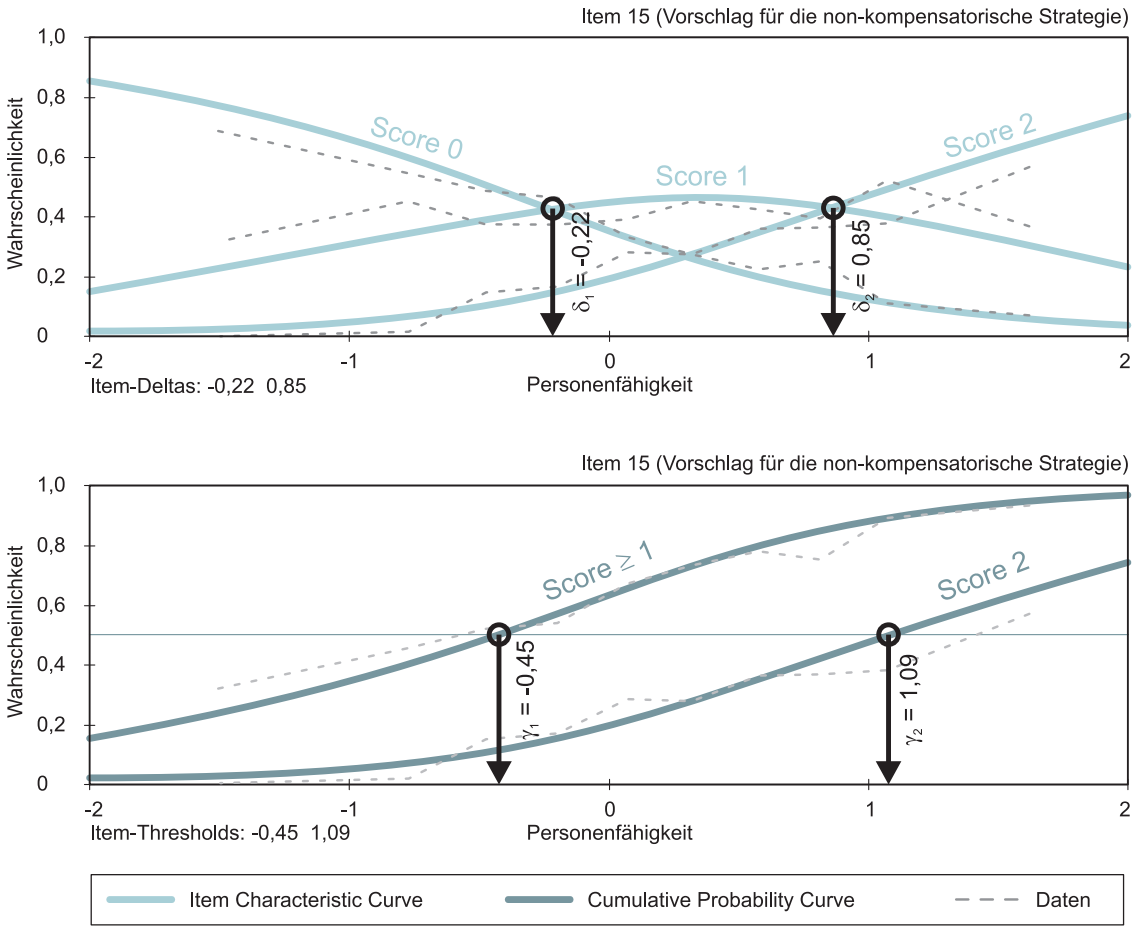

Abbildung 5.2: Unterschied zwischen Item-Deltas und Item-Thresholds. Oben: Item Characteristic Curve mit Schwellenparameter δ, der die Schwellen zur Antwortkategorie mit der jeweils höchsten Wahrscheinlichkeit kennzeichnet. Unten: Cumulative Probability Curve mit Schwellenparameter γ, der kennzeichnet, ab welcher Personenfähigkeit ein bestimmter Score oder ein höherer wahrscheinlicher ist (kumulierte Wahrscheinlichkeit).

Allgemeine Verfahrensempfehlung: Dem Grundsatz folgend, dass Kategorien mit höherem Score auch mit höheren Personenfähigkeiten korrespondieren sollten, sollten die Item-Thresholds ebenso innerhalb des Intervalls [-3; 3] liegen und geordnet sein: Sukzessiv ansteigende Thresholds zu höheren Scores.

Konkrete Handhabung in dieser Studie: Dies war ein Hauptkriterium, nach dem bei der Zusammenfassung der Scoring-Kategorien vorgegangen wurde. Immer kombiniert mit inhaltlichen Überlegungen wurden die Zusammenlegungen durch systematische Variationen explorativ so gewählt (siehe Tabelle 5.7), dass die Item-Thresholds möglichst im Intervall [-3; 3] liegen und geordnet sind, also zu höheren Scores ansteigen. Auffällig hinsichtlich dieses Kriteriums werden die Itemsteps 16.2 und 17.2, deren Item-Thresholds mit 3,01 bzw. 3,35 geringfügig außerhalb des anvisierten Intervalls von [-3; 3] liegen.

Average ability of students responding in each category (PV1Avg:1, ConQuest)

Im Gegensatz zur Pt Bis (die eine Korrelation des einzelnen Itemscores zum Gesamtscore herstellt) gibt der Parameter *PV1Avg:1* die mittlere Personenfähigkeit derjenigen Schülerinnen und Schüler an, die die jeweilige Antwortmöglichkeit gewählt haben (OECD 2009: S. 148; M. Wu, pers. Mitt. 23. September 2013).
Allgemeine Verfahrensempfehlung: Der PV1Avg:1-Wert sollte für falsche Kategorien geringer sein als für richtige Kategorien und demzufolge geordnet sein, also zu höheren Scores ansteigen (M. Wu, pers. Mitt. 23. September 2013).
Konkrete Handhabung in dieser Studie: Bei der Zusammenlegung von Kategorien wurde darauf geachtet, dass dieser Wert geordnet ist, also sukzessive ansteigt. Dieses Kriterium war weitestgehend unproblematisch, bis auf wenige der theoretisch möglichen Kategorien-Zusammenlegungen (siehe Tabelle 5.7) wurde diese Bedingung erfüllt. Die mittleren Personenfähigkeiten von Schülerinnen und Schülern, deren Antwort mit dem Score 0 versehen wurden, sind durchgehend kleiner als 0.

Zwischenfazit: Zusammenfassende Bemerkungen zur Itemselektion

Eine ausführliche Vorstellung der Item-Fit-Indizes aus den verbleibenden Items erfolgt in Kapitel 6.2 und eine Diskussion in Kapitel 7.2. An dieser Stelle sollen bereits ein paar Ergebnisse berichtet werden, um ein Verständnis der folgenden Unterkapitel über die ausgeschlossenen Items (siehe Kapitel 5.3.3) und über die Zusammenlegung von Kategorien (siehe Kapitel 5.3.4) zu ermöglichen.

Ausschluss von Items

In der Gesamtwürdigung wurde deshalb entschieden, geringe Auffälligkeiten von Items zur Kenntnis zu nehmen und zu überdenken. Drei Items fielen allerdings hinsichtlich mehrerer Item-Fit-Indizes auf (siehe Tabelle 5.5), weshalb sie von der Modellierung mit dem Rasch-Partial-Credit-Modell ausgeschlossen wurden und diese nur mit den in Tabelle 6.7 aufgeführten Items durchgeführt wurde.

geeignete Items

Grundsätzlich scheint die Mehrheit der Items den aufgeführten Item-Fit-Indizes auf Anhieb zu entsprechen. Bei 3 Items treten gehäuft Auffälligkeiten hinsichtlich einzelner Item-Fit-Indizes auf, weswegen sie aus der Modellierung ausgeschlossen wurden (siehe Kapitel 5.3.3 und Tabelle 5.5).

Overfits

Bei den Z_Q-Werten wurden Overfits der Items 4, 6 und 7 festgestellt (siehe Tabelle 6.7). Beim direkten Vergleich der Z_Q- und T-Werte zeigt sich allerdings, dass die Auffälligkeiten beim Z_Q-Index sich nicht im T-Wert widerspiegeln – mit Ausnahme des Items 4, das bei

Tabelle 5.5: Übersicht über die ausgeschlossenen Items.

Item-Fit-Parameter				
	Itemnummer:	X1	X2	X3
	Kürzel[a]	NeuA	NeuA	NeuA
Parameter[b]	Vorgabe	Aufgabe 1	Aufgabe 2	Aufgabe 3
Item-Fit-Parameter mit WINMIRA (N=718 komplette Datensätze)				
Q-Index	$< 0{,}35$	0,44[Q]	0,37[Q]	0,53[Q]
Z_Q	$\in [-1{,}96; 1{,}96]$	3,94[U]	1,54	2,39[U]
Z_Q	signifikant?	***	–	**
Thresholds	$\in [-3; 3]$	2,23	2,62	4,55[S]
Item-Fit-Parameter mit ConQuest (N variiert)				
N		848	849	850
WMNSQ	$\in [0{,}75; 1{,}33]$	1,05	1,00	0,99
dazu T-Wert	$\in [-2; 2]$	0,4	0,1	0,1
Trennschärfe	>0,25	0,05[T]	0,10[T]	-0,02[T]
Item-Deltas[c]	$\in [-3; 3]$	2,88	4,04[S]	5,41[S]
PV1Avg:1	<0 für Score 0	0,00	0,00	0,01
PV1Avg:1	>0 für Score 1	0,16	0,34	0,03

[a] NeuA: Schülerin oder Schüler nennt neue Aspekte
[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.
[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.
[Q] überschreitet maximal zulässigen Q-Index
[T] erreicht Mindest-Trennschärfe nicht
[U] Underfit eines Items
[S] Item ist zu schwierig, da WINMIRA-Threshold bzw. Item-Delta >3,00.

beiden Parametern einen Overfit ausweist. Es erscheint jedoch bei den Entscheidungsaufgaben nicht sinnvoll, einzelne Optionen aus dem Gesamtzusammenhang herauszulösen, zudem ist ein Item-Overfit nicht so streng zu bewerten wie ein Item-Underfits (Bühner 2006, S. 366, Bond und Fox 2010, S. 240).

Ebenso soll bei der Reflexionsaufgabe ein konsistentes Bild mit je einer Aufgabe zur Beschreibung und einer zum Geben von Verbesserungsvorschlägen erhalten bleiben. Die nicht erfüllten Anforderung des Z_Q-Wertes beim Item 13 werden daher zwar beobachtet, es wird aber nicht aus dem Itempool und der Rasch-Modellierung herausgehalten, zumal auch der T-Wert nicht auffällig wird (siehe Tabelle 6.7).

5.3.3 Ausgeschlossene Items: „Nennung neuer Aspekte“

Bei beiden Entscheidungsaufgaben und der Reflexionsaufgabe wurde zudem mit erfasst, ob Schülerinnen und Schüler neue, im Testheft nicht genannte, Aspekte in ihre Entscheidung

mit einfließen lassen. Die Codierung sah vor, hierfür den Score 1 zu vergeben und den Score 0, falls die Schülerinnen und Schüler keine neuen Aspekte in ihre Antwort einfließen lassen. Eine solche Codierung erfolgte in den Studien von Eggert und Bögeholz (2006, 2010) bisher nicht und hat sich auch als nicht zielführend für die Analyse im Rasch-Partial-Credit-Modell erwiesen: Diese Items konnten mehrere Fitparameter nicht erfüllen (siehe Tabelle 5.5).

5.3.4 Zusammenlegung von Kategorien

inhaltliche Überlegungen

Das Zusammenlegen der Kategorien und die damit verbundene Anpassung des Scoring Guides (siehe Kapitel 5.1.1 und 5.1.2 sowie Tabellen 5.1 und 5.2) erfolgte nach inhaltlichen Überlegungen, die dann hinsichtlich ihrer Auswirkung auf die Modellgüte innerhalb des Rasch-Partial-Credit-Modells überprüft wurden. Dabei stellt die Zusammenlegung von Kategorien nicht zwangsläufig eine Verbesserung der Modellpassung dar (Andrich et al. 1997: S. 61).

Beim Zusammenlegen von Kategorien geben auch die Werte für Pt Bis und PV1Avg:1 einen Hinweis darauf, ob die betreffende Kategorie mit der jeweils nächsthöheren oder nächstniedrigeren zusammengelegt werden sollte (Wu & Adams 2007: S. 68).

nur Nachbarkategorien betrachten

Dieser Prozess ist aufwendig, da die Auswirkungen jeder einzelnen Zusammenlegung von Kategorien auf die Modellierung unmittelbar überprüft werden muss. Grundsätzlich ist nur das Zusammenlegen von benachbarten Kategorien sinnvoll, bei m Kategorien sind dies also $m-1$ mögliche Zusammenlegungen pro Item. Doch auch mit Berücksichtigung dieser Einschränkung sind im Scoring Guide in dieser Studie rechnerisch noch eine Vielzahl von Zusammenlegungen bei den Items 1-5, 6-10 und 15-17 möglich. Auch die Häufigkeit, mit der bestimmte Kategorien angewählt werden, kann bei der Auswahl geeigneter Zusammenlegungen herangezogen werden (Wu & Adams 2007: S. 68), wovon in dieser Studie allerdings nur bedingt Gebrauch gemacht werden soll: Sinnvoll erscheint vielmehr, bei ähnlichen Aufgaben ganzheitlich vorzugehen und auch dieselben Kategorienzusammenfassungen anzuwenden, so dass z. B. in den Aufgaben 1 und 2 nur die Alternativen bleiben bei den Items 1-4 und 6-9 („Umgang mit Optionen“) die Kategorien (0 und 1) oder (1 und 2) zusammen zu legen, ebenso sollte für die Items 5 und 10 („Gewichtung von Kriterien“) eine einheitliche Zusammenlegung gefunden werden. Die zu überprüfenden Zusammenlegungen reduzieren sich damit zwar, deren Zahl ist aber immer noch zu hoch, um alle Kombinationen systematisch und unabhängig voneinander vorzunehmen. Es wurde daher in zwei Schritten vorgegangen:

Überprüfung der Kategorien der Entscheidungsaufgaben

Aufgaben 1 und 2

Zunächst wurden Zusammenlegungen bei den ersten Aufgaben (Items 1-10) analysiert und geprüft, wie sich einheitliche Zusammenlegungen bei den Items 1-4 und 6-9 („Umgang mit Optionen“) auf der einen und bei den Items 5 und 10 („Gewichtung von Kriterien“) auf der anderen Seite auswirken. Unter Anwendung inhaltlicher Überlegungen und als Resultat des Analyseprozesses wurde die Entscheidung getroffen, bei den Items 1-4 und 6-9 die ursprünglichen Scores 0 („Schüler/-in erwähnt keine Kriterien“) und 1 („Schüler/-in erwähnten nur positive oder nur negative Kriterien“) zusammenzulegen (siehe Tabellen 5.1 und 5.6). Auch bei der Gewichtung von Kriterien (Items 5 und 10) wurde im Rahmen der Raschmodellierung nur noch danach unterschieden ob und nicht wie viele Gewichtungen vorgenommen wurden.

Tabelle 5.6: Vorgehen bei der Überprüfung von Kategorienzusammenlegungen in den Entscheidungsaufgaben für die Optionen (Items 1-4, 6-9) und für die Gewichtungen von Kriterien (Items 5 und 10). Grau hinterlegte Zellen zeigen die jeweils überprüfte Zusammenlegung von Kategorien an.

Vorgehen bei Überprüfung von Kategorienzusammenlegungen bei den Entscheidungsaufgaben							
	Items 1-4, 6-9			Item 5, 10			
ursprüngliche Scores:	0	1	2	0	1	2	
angepasste Scores Variation I:	0		1	0	1	2	ausgewählt
angepasste Scores Variation II:	0	1		0	1	2	
angepasste Scores Variation I:	0	1	2	0		1	
angepasste Scores Variation II:	0	1	2	0	1		ausgewählt
Ergebnis:	0		1	0	1		

Überprüfung der Kategorien der Reflexionsaufgabe

Aufgabe 3

Mit diesen bereits neu eingeteilten Kategorien wurden dann die Auswirkungen von Zusammenlegungen in der Reflexionsaufgabe untersucht, wobei zunächst itemweise vorgegangen wurde (siehe Tabelle 5.7). Dies bedeutet, dass nur jeweils eine Variation von Scoring-Kategorien auf einmal vorgenommen wurde. Somit können die Auswirkungen einer Zusammenlegung auf das Gesamtmodell unabhängig von denen bei den anderen Items festgestellt werden. Sind diese dann bei den verschiedenen Items bekannt und werden sie durch die Modellierung gestützt, können die bestpassenden im Hinblick auf Gemeinsamkeiten gegenübergestellt werden. Wünschenswert wäre auch hier, wenn sich eine einheitliche Zusammenlegung der Kategorien in den Items 15 bis 17 realisieren ließe.

Bei Item 16 wurde ein alternatives Scoring getestet (identisch mit dem in Eggert & Bögeholz 2010), allerdings wurde sich aus inhaltlichen Überlegungen heraus (nämlich dass das

Anerkennen der geschilderten Entscheidungsstrategie als „schon ganz gut“ einen teilweise richtigen strategischen Vorschlag darstellt und daher nicht mit der Nullkategorie zusammengelegt werden sollte) dafür entschieden, nicht auf dieses Scoring zurückzugreifen.

Bei der Gesamtbetrachtung geeigneter Zusammenlegungen von Kategorien kann abschließend festgestellt werden, dass diese für das Item 17 („Verbesserungsvorschlag für das intuitiv-rechtfertigende Vorgehen“) aufgrund der Resultate aus der Modellierung anders ausfallen muss als für die Items 15 („Verbesserungsvorschlag für die non-kompensatorische Strategie“) und 16 („Verbesserungsvorschlag für die kompensatorische Strategie“).

Tabelle 5.7: Vorgehen bei Überprüfung von Kategorienzusammenlegungen bei den Items 15-17 („Verbesserungsvorschläge zu den Strategien“). Grau hinterlegte Zellen zeigen die jeweils überprüfte Zusammenlegung von Scoring-Kategorien an. Bei der Variation der Scoring-Kategorien zu einer jeden Aufgabe wurden die der anderen zunächst jeweils in ihrem ursprünglichen Scoring belassen.

Vorgehen bei Überprüfung von Kategorienzusammenlegungen bei den Items 15-17																	
	Item 15				Item 16					Item 16alt[a]				Item 17			
ursprüngliche Scores:	0	1	2	3	0	1	2	3	4	02	1	3	4	0	1	2	3	
ang. Scores Var. I:	0		1	2	0	1	2	3	4					0	1	2	3	
ang. Scores Var. II:	0	1		2	0	1	2	3	4					0	1	2	3	ausgewählt
ang. Scores Var. III:	0	1	2		0	1	2	3	4					0	1	2	3	
ang. Scores Var. I:	0	1	2	3	0		1	2	3					0	1	2	3	
ang. Scores Var. II:	0	1	2	3	0	1		2	3					0	1	2	3	
ang. Scores Var. III:	0	1	2	3	0	1	2		3					0	1	2	3	
ang. Scores Var. IV:	0	1	2	3	0	1	2	3						0	1	2	3	
ang. Scores Var. V:	0	1	2	3		0		1	2					0	1	2	3	
ang. Scores Var. VI:	0	1	2	3	0		1		2					0	1	2	3	ausgewählt
ang. Scores Var. VII:	0	1	2	3	0	1		2						0	1	2	3	
ang. Scores Var. VIII:	0	1	2	3						0		1	2	0	1	2	3	
ang. Scores Var. IX:	0	1	2	3						0	1		2	0	1	2	3	
ang. Scores Var. X:	0	1	2	3						0	1	2		0	1	2	3	
ang. Scores Var. I:	0	1	2	3	0	1	2	3	4						0	1	2	ausgewählt
ang. Scores Var. II:	0	1	2	3	0	1	2	3	4					0	1	2		
ang. Scores Var. III:	0	1	2	3	0	1	2	3	4					0	1	2		
Ergebnis:	0	1		2	0	1		2						0	1	2		

[a] Hierbei handelt es sich um ein alternatives Scoring, bei dem die ursprünglichen Kategorien 0 und 2 (siehe Tabelle 5.2) bereits zusammengefasst wurden. Dieses Scoring ist das von (Eggert & Bögeholz 2010) vorgeschlagene.

KAPITEL 6

Ergebnisse

6.1 Forschungsfrage I – Messbarkeit von *Bewerten, Entscheiden und Reflektieren*

6.1.1 Deskriptive Befunde der Kategorienvergabe

Häufigkeiten

Die Häufigkeiten, mit denen Schülerinnen und Schüler bestimmte (ursprüngliche) Codes für ihre Antworten bekommen haben, sind in Tabelle 6.1 dargestellt. Betrachtet man jeweils die höchsten Kategorien, so fällt auf, dass die Items 5 und 10 („Gewichten von Kriterien“) mit einem relativen Anteil von 11% bzw. 5% die schwierigsten der Entscheidungsaufgaben sind, in der Reflexionsaufgabe – und im gesamten Testheft – sind das Geben eines vollständig strategischen Verbesserungsvorschlags für die kompensatorische Strategie (Item 16) mit einem relativen Anteil von 5% und das Geben eines vollständig strategischen Verbesserungsvorschlags für das intuitiv-rechtfertigende Entscheidungsverhalten (Item 17) mit einem relativen Anteil von nur 1% die schwierigsten Items (siehe Tabellen 6.1 und 6.2). Hingegen ist die Antwort auf die Frage, ob Sabines oder Bernds Entscheidung die passendere ist (Item 14), mit einem relativen Anteil korrekter Antworten von 88% das leichteste Item im gesamten Test. Dies deckt sich mit den im späteren Teil dieser Arbeit durchgeführten Analysen im Rasch-Partial-Credit-Modell (siehe Abbildung 6.2). Itemschwierigkeiten werden in der probabilistischen Analyse noch einmal auf andere Weise berechnet und sind eine zentrale Größe der Raschmodellierung (siehe Kapitel 5.3.1).

Tabelle 6.1: Deskriptive Statistik der Antworten in der ursprünglichen Codierung (siehe Tabellen 5.1 und 5.2).

Itemnr.	*Score 0*	*Score 1*	*Score 2*	*Score 3*	*Score 4*	*N*	*Trennschärfe*[a]
1	304 (36%)	231 (27%)	313 (37%)	–	–	848	0,37
2	305 (36%)	194 (23%)	349 (41%)	–	–	848	0,28
3	254 (30%)	199 (23%)	395 (47%)	–	–	848	0,40
4	264 (31%)	221 (26%)	363 (43%)	–	–	848	0,49
5	647 (76%)	111 (13%)	90 (11%)	–	–	848	0,13
6	414 (49%)	273 (32%)	162 (19%)	–	–	849	0,54
7	360 (42%)	230 (27%)	259 (31%)	–	–	849	0,49
8	265 (30%)	191 (22%)	393 (46%)	–	–	849	0,36
9	160 (19%)	242 (29%)	447 (53%)	–	–	849	0,33
10	699 (82%)	109 (13%)	41 (5%)	–	–	849	0,16
11	215 (25%)	633 (75%)	–	–	–	848	0,23
12	279 (33%)	565 (67%)	–	–	–	844	0,22
13	574 (68%)	269 (32%)	–	–	–	843	0,11
14	94 (12%)	707 (88%)	–	–	–	801	0,14
15	313 (37%)	127 (15%)	205 (25%)	191 (23%)	–	836	0,26
16	269 (34%)	33 (4%)	400 (51%)	42 (5%)	36 (5%)	780	0,15
17	455 (56%)	206 (26%)	134 (17%)	11 (1%)	–	806	0,15

Prozentuale Angaben beziehen auf gültige Antworten.

[a] Trennschärfe als korrigierte Item-Skala-Korrelation beim listenweisen Fallausschluss beziehen sich auf 718 vollständige Datensätze.

Tabelle 6.2: Deskriptive Statistik der Antworten in der angepassten Codierung aus Zusammenlegungen von Kategorien (siehe Tabellen 5.1 und 5.2).

Itemnr.	*Score 0*	*Score 1*	*Score 2*	*N*
1	535 (63%)	313 (37%)	–	848
2	499 (59%)	349 (41%)	–	848
3	453 (53%)	395 (47%)	–	848
4	485 (57%)	363 (43%)	–	848
5	647 (76%)	201 (24%)	–	848
6	687 (81%)	162 (19%)	–	849
7	590 (69%)	259 (31%)	–	849
8	456 (54%)	393 (46%)	–	849
9	402 (47%)	447 (53%)	–	849
10	699 (82%)	150 (18%)	–	849
11	215 (25%)	633 (75%)	–	848
12	279 (33%)	565 (67%)	–	844
13	574 (68%)	269 (32%)	–	843
14	94 (12%)	707 (88%)	–	801
15	313 (37%)	332 (40%)	191 (23%)	836
16	269 (34%)	475 (61%)	36 (5%)	780
17	661 (82%)	134 (17%)	11 (1%)	806

Prozentuale Angaben beziehen auf gültige Antworten.

6.1.2 Objektivität

Die Durchführungsobjektivität der Erhebung wird durch die Vorlage eines Manuals gesichert, in dem der Ablauf jeder Testsitzung genau beschrieben ist. Zudem wurden alle Testsitzungen von demselben Testleiter durchgeführt. In Bezug auf die Auswerteobjektivität wird an dieser Stelle auf die Scoring Guides in den Tabellen 5.1 und 5.2 und die Darstellung der Inter-Rater-Reliabilität in Kapitel 5.1.3 verwiesen.

6.1.3 Reliabilität und Interne Konsistenz

Trennschärfe

Als Maß für die *Interne Konsistenz* kann zum Beispiel die Korrelation eines Items mit der Gesamtskala des Tests herangezogen werden. Dabei wird die Skala jedoch aus verbleibenden Items gebildet, also ohne das Item, für das eine Trennschärfe berechnet wird („part-whole-Korrektur"). Diese *Trennschärfen* der Items sind ebenfalls in der Tabelle 6.1 angegeben, allerdings beziehen sie sich auf vollständige Datensätze ($N = 718$) und die ursprüngliche Codierung. Dabei fällt auf, dass einige Items sehr geringe Trennschärfen im Bereich 0,11-0,16 aufweisen (Items 5, 10, 13, 14, 16 und 17), andere Items hingegen vergleichsweise hohe Trennschärfen im Bereich 0,49-0,54 (Items 4, 6 und 7). Eine Trennschärfe von mindestens 0,25 bis 0,30 wird gelegentlich in der Literatur gefordert; feste Untergrenzen, ab denen Items ausgeschlossen werden sollten, sind als alleiniges Ausschlusskriterium jedoch nicht zielführend (Bühner 2012: S. 244,246f). Zudem werden sich im Rahmen der Scorezusammenlegungen (siehe Tabellen 5.1 und 5.2) für die probabilistische Auswertung noch andere Trennschärfen ergeben, die in Tabelle 6.7 aufgeführt sind.

Cronbachs $\alpha = 0{,}69$ bzw. $\alpha = 0{,}65$

Auf Basis der vollständigen Datensätze ($N = 718$) in der ursprünglichen Codierung (siehe Tabellen 5.1 und 5.2) kann ein Cronbach-Alpha angegeben werden mit $\alpha_{ursprünglich} = 0{,}69$[1] und in der für die probabilistischen Testtheorie (siehe Kapitel 5.3) angepassten Codierung $\alpha_{angepasst} = 0{,}65$[2], was eine hinreichende innere Konsistenz gewährleistet. Die mit ConQuest probabilistisch ermittelte WLE-Reliabilität im angepassten Scoring beträgt $0{,}64$[3] ($N = 850$).

1 Die fünf mittels multipler Imputation vervollständigten Datensätze liefern jeweils $\alpha_{ursprünglich} = 0{,}71$.
2 Die fünf mittels multipler Imputation vervollständigten Datensätze liefern jeweils $\alpha_{angepasst} = 0{,}66$.
3 Die fünf mittels multipler Imputation vervollständigten Datensätze liefern ebenfalls WLE-Reliabilitäten von 0,63-0,64.

6.1.4 Konstruktvalidität

Bewerten, Entscheiden und Reflektieren und Schulalter

Tabelle 6.3: Mittelwerte hinsichtlich der WLE-Schätzer für *Bewerten, Entscheiden und Reflektieren* hinsichtlich der einzelnen Subgruppen. Die Mittelwerte sind auch in Abbildung 6.1 dargestellt.

Subgruppe	*Mittelwert* $\theta_{gem.}$	*Standardabweichung*
Klasse 6	-0,81	0,87
Klasse 8	0,04	0,80
Klasse 10	0,31	0,76
Oberstufe	0,53	0,73
Gesamt	0,00	0,94

eindeutiger Zusammenhang

Die Ergebnisse der Datenanalyse zeigen einen eindeutigen Zusammenhang von Schulalter und *Bewerten, Entscheiden und Reflektieren*. Zur Überprüfung auf signifikante Mittelwertunterschiede in den einzelnen Altersklassen wurde eine einfaktorielle Varianzanalyse (ANOVA) mit geplanten Kontrasten (jeweils zwischen den Klassen 6 und 8, 8 und 10, 10 und Oberstufe) durchgeführt. Die Varianzen können über die Altersgruppen hinweg als homogen angesehen werden[1]. Weitere paarweise Tests auf Varianzhomogenität zwischen den jeweils an den geplanten Kontrasten beteiligten Subgruppen unterstützen diese Aussage.

Wie in Tabelle 6.4 ersichtlich, können für alle Kontraste signifikante Unterschiede in den jeweiligen Altersgruppen festgestellt werden: Schülerinnen und Schüler der jeweils höheren Altersgruppe erreichen signifikant bessere Ergebnisse als Schülerinnen und Schüler der jeweils niedrigeren Altersstufe.

Tabelle 6.4: Kontraste hinsichtlich des WLE-Schätzers für *Bewerten, Entscheiden und Reflektieren* in den verschiedenen Altersgruppen.

Kontrast	*Kontrastwert*[a]	*Standardfehler*	*Signifikanz*
Klasse 6 zu 8	+0,84	0,08	p<.001
Klasse 8 zu 10	+0,28	0,07	p<.001
Klasse 10 zu Oberstufe	+0,22	0,08	p=.005

[a] Differenz der Mittelwerte in den jeweiligen Gruppen

Die ebenfalls durchgeführte Berechnung einer Korrelation[2] zwischen den Variablen „Schulalter“ und „WLE-Schätzer“ liefert eine alternative Angabe dieses Zusammenhangs $r = 0{,}51^{**}$

1 Der Levene-Test ergibt F=1,27 bei df1=3 und df2=846 bei der gesamten Stichprobe und bleibt damit deutlich unter dem kritischen F-Wert von 2,61 für $F(3,\infty)$, ab dem von inhomogenen Varianzen auszugehen ist. Analoge Formulierung: Der Levene-Test wird nicht signifikant (p=0.28).

2 Spearman-Korrelation für ordinalskalierte Daten

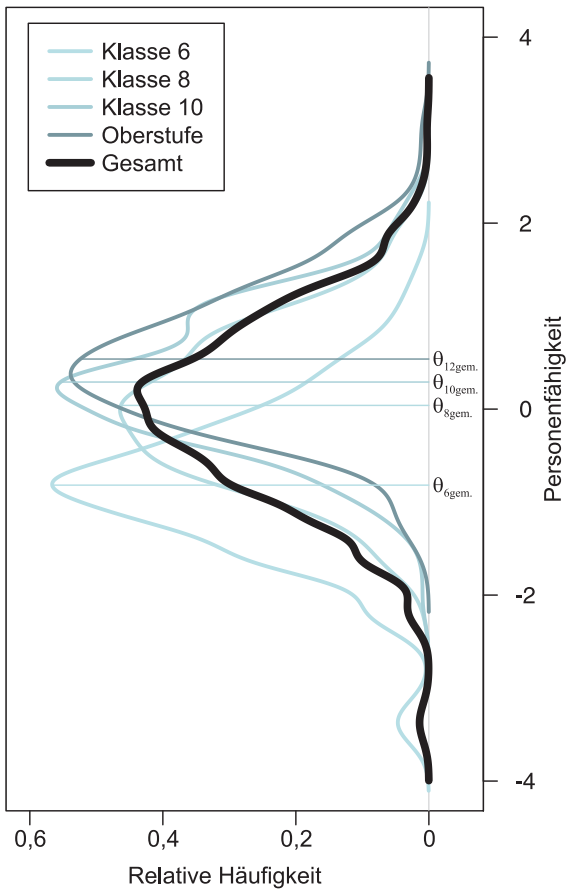

Abbildung 6.1: Vergleich der Verteilungen für *Bewerten, Entscheiden und Reflektieren* innerhalb der verschiedenen Jahrgangsgruppen, die sich hinsichtlich ihres Mittelwerts signifikant unterscheiden (siehe Tabelle 6.4).

und weist damit einen starken Einfluss ($r^2 = 0{,}26$) des Schulalters auf *Bewerten, Entscheiden und Reflektieren* aus.

Bewerten, Entscheiden und Reflektieren und Schulnoten

Zusammenhang mit manchen Schulnoten

Zu den erhobenen Schulnoten für die Unterrichtsfächer Biologie, Deutsch, Englisch, Erdkunde, Mathematik, Physik und Politik wurden Korrelationen zu *Bewerten, Entscheiden und Reflektieren* berechnet und damit Rückschlüsse auf die Konstruktvalidität des Testinstruments gezogen. Die Korrelationen wurden dabei sowohl einzeln für jeden Jahrgang als auch aggregiert für die Klassen 6 bis 10 berechnet. Die Subgruppe *Oberstufe* wurde in diese Korrelation nicht mit eingerechnet, da im Kurssystem der Oberstufe Wahleffekte auftreten können, wie man unter anderem an den unterschiedlichen Fallzahlen in Tabelle 6.5 sieht. Das Unterrichtsfach Physik ist kein Pflichtfach mehr und es ist nicht auszuschließen, dass

es nur von interessierten Schülerinnen und Schülern angewählt wird. Zudem sind die Stundenanzahlen der Fächer nicht mehr miteinander zu vergleichen, da sie je nach Wahl der Schülerinnen und Schüler zwei- oder vierstündig unterrichtet werden.

Tabelle 6.5: Korrelationen von *Bewerten, Entscheiden und Reflektieren* mit ausgewählten Schulnoten.

Korrelationen von Bewerten, Entscheiden und Reflektieren mit ausgewählten Schulnoten										
	Klasse 6		*Klasse 8*		*Klasse 10*		*Oberstufe*		*Z(Klassen 6-10)*[a]	
Schulfach	r[b]	N	r[b]	N	r[b]	N	r[b]	N	r[b]	N
Biologie	0,24***	189	0,18**	213	0,12	212	0,19*	147	0,21***	614
Deutsch	0,09	199	0,29***	221	0,26***	216	0,10	169	0,28***	636
Englisch	0,10	201	0,32***	218	0,13	217	0,09	163	0,22***	636
Erdkunde	0,17*	196	0,27***	214	0,11	211	0,00	91	0,23***	621
Mathematik	0,11	201	0,31***	220	0,27***	217	0,04	172	0,22***	638
Physik	0,01	183	0,23**	217	0,14*	218	0,32**	95	0,15**	618
Politik	0,78	4	0,24	12	0,14*	212	0,09	156	0,16*	228

[a] jeweils jahrgangsweise z-standardisiert
[b] Spearman-Korrelation für Ordinalskalierung. Signifikanzniveaus: *α=0,05 **α=0,01 ***α=0,001

Die berechneten Korrelationen sind in Tabelle 6.5 zusammengestellt. Grundsätzlich sollten nur die signifikanten Korrelationen betrachtet und ausgewertet werden (gekennzeichnet durch mindestens ein Sternchen). Bei der Analyse der berechneten Zusammenhänge lässt sich feststellen, dass insgesamt nur geringe Korrelationen zu Schulnoten auftreten. Unter ihnen befinden sich die höchsten Korrelationen in der 8. Klasse, sie gehen in den höheren Jahrgängen wieder zurück. In der Klasse 10, die mit dem Sekundarabschluss I endet und somit hinsichtlich der entworfenen Bildungsstandards besonders interessant ist, sind nur Korrelationen zur Deutsch-, Mathematik-, Physik- und Politiknote signifikant. Für die Oberstufe sind höhere Korrelationen im Hinblick auf die Biologie- und Physiknote festzustellen, alle anderen Korrelationen fallen niedriger aus und sind damit auch nicht mehr signifikant.

Betrachtet man alle Klassen der Sekundarstufe I (Klasse 6 bis 10) aggregiert, so lassen sich aufgrund der höheren Fallzahlen wieder signifikante Korrelationen angeben (siehe Tabelle 6.5, rechts). Dazu wurden die Schulnoten jahrgangsweise z-standardisiert und mit den ebenfalls jahrgangsweise z-standardisierten WLE-Personenfähigkeiten korreliert, um den gegeneinander verschobenen Verteilungskurven in den verschiedenen Jahrgängen gerecht zu werden. Die Korrelationen zu allen Schulfächern werden signifikant und lassen sich damit untereinander vergleichen. Die größte Korrelation ergibt sich mit $r = 0{,}28^{***}$ zur Deutschnote, die geringste Korrelation mit $r = 0{,}15^{**}$ zur Physiknote.

Zusammenhang von *Bewerten, Entscheiden und Reflektieren* mit Leseverständnis und -geschwindigkeit

Nicht unwichtig ist, inwieweit das gemessene Testergebnis durch Lesekompetenz beeinflusst wird. Um *Bewerten, Entscheiden und Reflektieren* im Hinblick auf Lesekompetenz abzugrenzen, wurde diese begleitend mittels des *Lesegeschwindigkeits- und -verständnistests (LGVT)* (Schneider et al. 2007) erhoben. Die in diesem Test erzielten Rohpunkte der Teilskalen *Lesegeschwindigkeit* und *Leseverständnis* der Probanden können mit dem Testergebnis für *Bewerten, Entscheiden und Reflektieren* in Beziehung gesetzt werden. Dabei wird das „Alter“ der Schülerinnen und Schüler (erhoben als Klassenstufe der Probanden) als Kontrollvariable durch die Verwendung einer partiellen Korrelation herausgerechnet (Zöfel 2003: S. 164ff), da dieses einen Zusammenhang mit *Bewerten, Entscheiden und Reflektieren* aufweist (siehe Kapitel 6.1.4).

Für den Zusammenhang zwischen Lesegeschwindigkeit und Personenfähigkeit für *Bewerten, Entscheiden und Reflektieren* wurde eine (partielle) Korrelation von $r < 0{,}01$ (p=0,90) gefunden. Der Zusammenhang zwischen Leseverständnis und Personenfähigkeit fällt mit einer (partiellen) Korrelation signifikant aus, ist aber mit $r = 0{,}12^{***}$ immer noch sehr gering (siehe Tabelle 6.6).

Tabelle 6.6: Korrelationen von *Bewerten, Entscheiden und Reflektieren* mit Lesegeschwindigkeit und Leseverständnis in den einzelnen Jahrgangsstufen.

Subgruppe	*Lesegeschwindigkeit*		*Leseverständnis*		*N*
	r	p-Wert[a]	r	p-Wert	
Klasse 6	0,02	0,75	0,14	0,04	219
Klasse 8	-0,12	0,06	0,05	0,44	229
Klasse 10	0,06	0,39	0,12	0,09	222
Oberstufe	0,10	0,19	0,12	0,10	180
Gesamt[a]	<0,01	0,90	0,12	<0,001	850

[a] hier als partielle Korrelation unter der Kontrollvariable *Schulalter*

6.2 Forschungsfrage II – Modellierung

6.2.1 Item-Fit-Parameter und Zusammenlegungen von Kategorien

Es konnten inhaltlich begründete Zusammenlegungen von Kategorien identifiziert werden, die die Modellierung von der Daten innerhalb eines eindimensionalen Rasch-Partial-Credit-Modells hinsichtlich der in Kapitel 5.3.2 beschriebenen Item-Fit-Indizes ermöglichen. Diese

Tabelle 6.7: Item-Fit-Indizes des Original-Datensatzes mit den Programmen WINMIRA und ConQuest beim eindimensionalen Rasch-Partial-Credit-Modell.

Item-Fit-Parameter																		
	Itemnummer:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Kürzel[a]	Opt	Opt	Opt	Opt	Gew	Opt	Opt	Opt	Opt	Gew	MCi	MCn	MCk	PE	VSn	VSk	VSi
Parameter[b]	Vorgabe	Aufgabe 1					Aufgabe 2					Aufgabe 3						
Item-Fit-Indizes mit WINMIRA (N=718 komplette Datensätze)																		
Q-Index	$< 0{,}35$	0,22	0,25	0,21	0,17	0,30	0,18	0,20	0,25	0,23	0,32	0,26	0,26	0,32	0,31	0,24	0,30	0,26
Z_Q	$\in [-1{,}96; 1{,}96]$	-1,03	0,04	-1,47	-3,00^O	1,35	-3,06^O	-2,31^O	0,05	-0,61	1,83	-0,22	-0,14	2,51^U	0,27	1,82	1,73	0,38
Z_Q	signifikant?	–	–	–	**	–	**	*	–	–	*	–	–	**	–	*	*	–
Thresholds	$\in [-3; 3]$	0,17	0,02	-0,29	-0,13	0,80	1,20	0,54	-0,29	-0,56	1,21	-1,72	-1,27	0,42	-2,62	-0,12	0,69	1,95
Item-Fit-Indizes mit ConQuest (N variiert)																		
N		848	848	848	848	848	849	849	849	849	849	848	844	843	801	836	780	806
WMNSQ	$\in [0{,}75; 1{,}33]$	0,96	1,00	0,95	0,91	1,01	0,93	0,97	0,98	0,98	1,03	0,97	0,99	1,07	1,03	1,06	1,06	0,99
dazu T-Wert	$\in [-2; 2]$	-1,3	0,1	-2,1^O	-4,1^O	0,3	-1,4	-0,9	-0,7	-1,0	0,6	-0,8	-0,4	2,0	0,4	1,6	1,3	-0,2
Trennschärfe	$>0{,}25$	0,44	0,40	0,50	0,54	0,30	0,45	0,45	0,44	0,46	0,26	0,39	0,37	0,28	0,26	0,48	0,33	0,34
Item-Deltas[c]	$\in [-3; 3]$	0,61	0,41	0,16	0,33	1,31	1,61	0,93	0,17	-0,12	1,71	-1,21	-0,79	0,86	-2,20	0,32	1,16	2,47

Item-Fit-Indizes für polytome Items																
		WINMIRA Thresholds für Score...			ConQuest Item-Thresholds für Score...			Item-Deltas für Score...			Pt Bis[d] für Score...			PV1Avg:1 für Score...		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Item 15	VSn	–	-0,64	0,41	–	-0,45	1,09	–	-0,22	0,85	-0,40	0,05	0,40	-0,29	0,05	0,44
Item 16	VSk	–	-1,10	2,47	–	-0,68	3,01^S	–	-0,66	2,98	-0,31	0,23	0,16^N	-0,26	0,13	0,52
Item 17	VSi	–	1,29	2,61	–	1,58	3,35^S	–	1,77	3,17^S	-0,34	0,31	0,13^N	-0,08	0,44	0,62

[a] Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtfertigendem Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtfertigendes Vorgehen, non-kompensatorischer und kompensatorischer Strategie

[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.

[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.

[d] Alle Pt Bis-Korrelationen sind signifikant mit $p<0{,}001$, bis auf die für Score 1 bei Item 15 (nicht signifikant).

[O, U] Leichter Overfit bzw. Underfit eines Items

[S] Item ist geringfügig zu schwierig, da Item-Threshold bzw. Item-Delta $>3{,}00$.

[N] Score 2 bei diesem Item ist nicht geordnet, aber positiv mit Gesamtskala korreliert.

sind in der Tabelle 6.7 aufgeführt, die Tabellen A.1 bis A.5 enthalten die Ergebnisse der Datenanalyse für die mittels multipler Imputation vervollständigten Datensätze.

Kleinere Abweichungen bei Items 4, 6 und 7

Die Item-Fit-Parameter in der Tabelle 6.7 sind größtenteils im Normbereich und zeigen, dass das gewählte Rasch-Partial-Credit-Modell insgesamt gut auf die empirischen Daten passt. Kleinere Abweichungen gibt es bei den Items 4, 6 und 7, hier ist beim Z_Q-Index ein leichter Overfit zu beobachten, diese Items passen also „zu gut“ zum gewählten Modell, allerdings liegen die Z_Q-Werte für Item 7 bei Betrachtung der mittels multipler Imputation vervollständigten Datensätze (siehe Tabellen A.1 bis A.5) innerhalb des anvisierten Intervalls und sind somit unauffällig. Beim Item 4 setzt sich diese Beobachtung im „verwandten“ T-Wert des WMNSQ-Index fort, bei den Items 6 und 7 (Optionen „Wasserstoff“ und „Druckluft“) verschwindet diese Auffälligkeit hier. Weiterhin ist ein Overfit bei Item 3 hinsichtlich des T-Werts zu beobachten, der sich beim Z_Q-Index des Original-Datensatzes noch nicht gezeigt hat, wohl aber bei der Analyse der mittels multipler Imputation vervollständigten Datensätze.

6.2.2 Person-Item-Map

Die *Person-Item-Map* zum *Bewerten, Entscheiden und Reflektieren* stellt das zentrale Ergebnis der Raschmodellierung dar. In ihr sind Personenfähigkeiten (links) und Itemschwierigkeiten (rechts) entlang einer gemeinsamen Achse aufgetragen. Die Personenfähigkeiten streuen um den Mittelwert $\mu = 0{,}00$ logits mit $\sigma = 0{,}94$ logits, jedes X repräsentiert 9,2 Personen. Die Abbildung 6.1 (schwarze Kurve) ergänzt diese und bildet den Verlauf der Verteilung zu den Extremen detaillierter nach.

Die Person-Item-Map lässt sich wie folgt interpretieren: Das Item 7 (Option „Druckluft“) mit einer Itemschwierigkeit (Item-Deltas bei ConQuest, siehe Kapitel 5.3.2) von 0,93 logits wird von Personen mit derselben Personenfähigkeit mit einer 50%igen Wahrscheinlichkeit gelöst, fähigere Personen haben eine höhere Wahrscheinlichkeit, das Item richtig zu lösen (und umgekehrt). Es ist dabei deutlich schwieriger als das Item 9 (Option „Bleiakku“) mit einer Itemschwierigkeit von -0,12 logits.

Die jeweiligen Itemschwierigkeiten sind als Zahlenwerte auch der Tabelle 6.7 als „Item-Deltas (ConQuest)“ zu entnehmen. Fasst man die Itemschwierigkeiten der drei polytomen Items bei den jeweiligen Strategien zusammen, so ergeben sich mittlere Itemschwierigkeiten für die non-kompensatorische Strategie von 0,32 logits, für die kompensatorische Strategie von 1,16 logits und für das intuitiv-rechtfertigende Entscheidungsverhalten von 2,47 logits (in Abbildung 6.2 sind jeweils die Schwierigkeiten der Itemsteps aufgetragen).

Lage der Items

Eine genauere Analyse der Lage von Itemschwierigkeiten der zweiten Entscheidungsaufgabe

Logits	Personen	Items Aufgabe 1 „Energieerzeugung“	Items Aufgabe 2 „Energiespeicherung“	Items Aufgabe 3 „Energienutzung“
3				16.2 VSk 17.2 VSi
2				
			10 Gew	
			6 Opt	17.1 VSi
	XX			
		5 Gew		
	XX			
				15.2 VSn
1	XXXX		7 Opt	
				13 MCk
	XXXXXX			
	X	1 Opt		
	XXXXXXX			
		2 Opt		
	XXXXXXX	4 Opt		
	XXXXXXXXXX			
	X	3 Opt	8 Opt	
0	XXXXXXXXX			
	X			
	XXXXXXXXXX		9 Opt	
	X			
	XXXXXXXX			
	X			15.1 VSn
	XXXXXXXXX			
				16.1 VSk
	XXXXXX			12 MCn
-1	XXXX			
	X			
	X			11 MCi
-2				
				14 PE
-3				

Abbildung 6.2: Person-Item-Map. Jedes X repräsentiert 9,2 Personen. Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtf. Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtf. Vorgehen, non-kompensatorischer und kompensatorischer Strategie; PE: Passendere Entscheidung.

in der Person-Item-Map für die Items 6-9 (Umgang mit Optionen) zeigt in Relation zur ersten die Besonderheit, dass diese Hierarchisierung über alle Jahrgänge hinweg konstant bleibt (siehe Abbildung 6.3): Die Diskussion der Option *Bleiakku* fällt Schülerinnen und Schülern aller Altersstufen am leichtesten, gefolgt von der Diskussion des *Pumpspeicherkraftwerks* und des *Druckluftspeichers*. Die Diskussion der Option *Wasserstoff* ist durchgängig die schwierigste.

Bei der Reflexionsaufgabe (und im gesamten Testinstrument) ist das Item, was die Schülerinnen und Schüler nach ihrer Meinung zur passenderen Entscheidung bei der Erfüllung der Wünsche der beiden Personen fragt (Item 14), das leichteste. Das zweitleichteste Item ist die Beschreibung des intuitiv-rechtfertigenden Entscheidungsverhaltens in Form einer MC-Aufgabe (Item 11), gefolgt von der Beschreibung der non-kompensatorischen Strategie (ebenfalls in Form einer MC-Aufgabe, Item 12). Die anspruchsvollste unter den MC-Aufgaben ist die Beschreibung der kompensatorischen Strategie (Item 13) mit einem logit-Wert von 0,86.

Das Geben eines inhaltlichen, teilweise richtigen strategischen Vorschlags für die kompensatorische Strategie oder die Feststellung, dass das Vorgehen „schon in Ordnung sei“ (allesamt mit Score 1 (siehe Tabelle 5.2) versehen, Itemstep 16.1) ist das drittleichteste Item, gefolgt vom Geben eines inhaltlichen oder teilweise richtigen strategischen Vorschlags für die non-kompensatorische Strategie (ebenfalls Score 1, Itemstep 15.1). Beim Verbesserungsvorschlag für das intuitiv-rechtfertigende Verhalten (Item 17) musste eine andere Zusammenlegung von Kategorien angewendet werden als bei den Items 15 und 16 (Verbesserungsvorschläge für die non-kompensatorische bzw. die kompensatorische

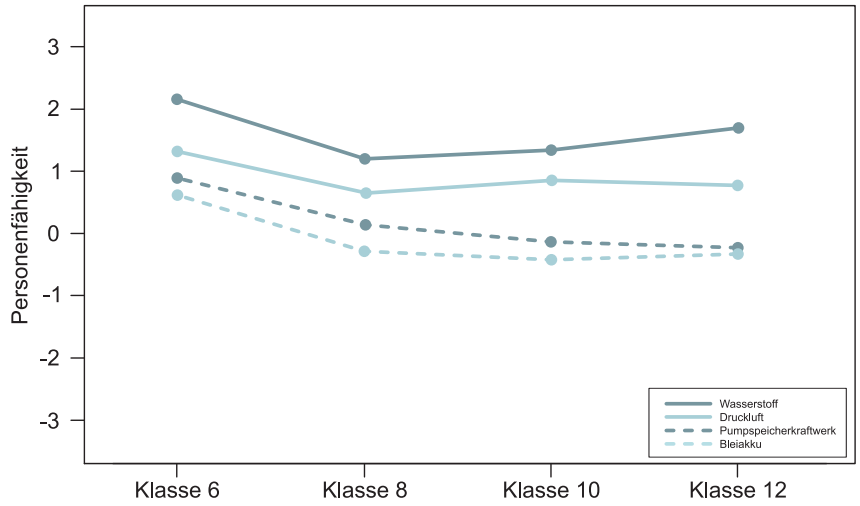

Abbildung 6.3: Verlauf der Itemschwierigkeiten der zweiten Entscheidungsaufgabe für die Items 6-9 (Umgang mit Optionen) in den einzelnen Klassenstufen.

Strategie; siehe Tabelle 5.2), um dieses Item in die Modellierung mit dem eindimensionalen Rasch-Partial-Credit-Modell mit einbeziehen zu können. Vermutlich resultiert hieraus auch die vergleichsweise hohe Schwierigkeit für dieses Item von 1,77 logits (Itemstep 17.1), das damit sogar noch schwieriger ist, als einen vollständig strategischen Vorschlag für die non-kompensatorische Strategie unterbreiten zu können (Itemstep 15.2). Die beiden mit Abstand schwierigsten Items im Testinstrument sind das Geben von vollständig strategischen Vorschlägen für die kompensatorische Strategie (Itemstep 16.2) mit einer Itemschwierigkeit von 2,98 logits und das Geben eines vollständig strategischen Verbesserungsvorschlags für das intuitiv-rechtfertigende Entscheidungsverhalten (Itemstep 17.2) mit einer Itemschwierigkeit von 3,17 logits.

6.3 Forschungsfrage III – Umgang mit Kriterien

6.3.1 Anzahl und Häufigkeiten verwendeter Kriterien in Entscheidungsprozessen

Kriterien unterschiedlich attraktiv

Die in Tabellenform den Schülerinnen und Schülern zur Verfügung gestellten Kriterien zu den Bearbeitungskontexten *Energiegewinnung* und *Energiespeicherung* scheinen in geringem Maße „unterschiedlich attraktiv“ für die Verwendung im Bewertungsprozess zu sein. Dabei ist das Kriterium *Energieverlust beim Transport* in der ersten Entscheidungsaufgabe (Energiegewinnung) das am wenigsten berücksichtigte, während das Kriterium zum *Naturschutz* vergleichsweise häufig in die Bewertung mit einbezogen wird. Interessanterweise zählen ökonomische Aspekte, wie z. B. die *Kosten für Bau und Wartung*, aber auch die *Windstärke*, erst in den Jahrgängen 10 und 12 zu den am häufigsten verwendeten Kriterien (siehe Tabelle 6.8 oben). Bei dieser Auflistung der Häufigkeiten verwendeter Kriterien ist zu erwähnen, dass sie nicht bei jeder Option genannt werden müssen: Es genügt das Aufführen eines Kriteriums bei mindestens einer Option, um es als verwendet zu zählen.

In der zweiten Entscheidungsaufgabe (Energiespeicherung) gehören die *Standortanforderungen* an die jeweilige Speichertechnologie zu den am wenigsten berücksichtigten Kriterien, während die *Speichergüte* (Wirkungsgrad) zu den am häufigsten verwendeten Kriterien gehört (siehe Tabelle 6.8 unten).

Für *Bewerten, Entscheiden und Reflektieren* im Kontext nachhaltiger Entwicklung im Themenfeld Energiegewinnung, -speicherung und -nutzung ist es insofern interessant hervorzuheben, dass die Kriterien *Windstärke* in der ersten Entscheidungsaufgabe und *Speichergüte* in der zweiten Entscheidungsaufgabe jeweils mit zu den am häufigsten verwendeten Kriterien in den Bewertungsprozesse gehören, als dass diese Kriterien auch Aspekte

der Ressourceneffizienz aufgreifen. Für das Kriterium *Energieverlust bei Transport* – das ebenfalls Aspekte der Ressourceneffizienz enthält – scheint bei Schülerinnen und Schülern aufgrund der geringen Verwendung allerdings vergleichsweise nur wenig Bewusstsein vorhanden zu sein.

Alters-abhängigkeit

Bei der Betrachtung hinsichtlich der verschiedenen Altersgruppen ist festzustellen, dass sich jeweils nur der Unterschied der Nennungen von der 6. Klasse zur 8. Klasse als signifikant erweist, allerdings mit der Ausnahme, dass sich das Kriterium *Sicherheitsrisiken* in allen untersuchten Jahrgangsstufen als durchweg gleichmäßig berücksichtigt herausstellt (siehe Tabelle 6.8).

Tabelle 6.8: Häufigkeiten der Verwendung dargebotener Kriterien in den Entscheidungsaufgaben, aufgeschlüsselt nach Jahrgängen der Schülerinnen und Schüler.

Häufigkeit der Verwendung verschiedener Kriterien bei Aufgabe 1 (Energieerzeugung)						
Subgruppe	*Windstärke*	*Naturschutz*	*Sichtbarkeit/Lärm*	*Energieverlust*	*Kosten*	*keine*
6	71%	81%	65%	60%	66%	1%
8	86%*	91%*	89%*	83%*	80%*	1%
10	88%	86%	85%	83%	86%	0%
12	92%	86%	86%	83%	90%	2%
Gesamt	84%	86%	81%	77%	80%	1%

Häufigkeit der Verwendung verschiedener Kriterien bei Aufgabe 2 (Energiespeicherung)					
Subgruppe	*Speichergüte*	*Standortanforderungen*	*Kosten/kWh*	*Sicherheitsrisiken*	*keine*
6	82%	78%	83%	87%	3%
8	93%*	89%*	91%*	89%	1%
10	91%	89%	92%	88%	0%
12	95%	90%	89%	88%	1%
Gesamt	90%	86%	89%	88%	1%

* Signifikanter Anstieg bzw. Abfall der Nennungen im Kontrast zu den Nennungen des jeweils vorausgehenden Jahrgangs

Anzahl einbezogener Kriterien

Auch hinsichtlich der Anzahl in den Bewertungsprozess einbezogener Kriterien lassen sich Unterschiede in den jeweiligen Jahrgangsgruppen feststellen. Mit zunehmendem Alter beziehen Schülerinnen und Schüler auch zunehmend mehr Kriterien in ihre Bewertung mit ein (siehe Abbildung 6.4). Bei der Untersuchung geplanter Kontraste fallen in den beiden Entscheidungsaufgaben allerdings jeweils auch hier nur wieder die Unterschiede zwischen der 6. Klasse und 8. Klasse signifikant aus[1], nicht aber die Unterschiede hinsichtlich der Anzahl verwendeter Kriterien zwischen Schülerinnen und Schülern der 8., 10. und 12. Klasse. Mit Ausnahme der 6. Klasse in der ersten Entscheidungsaufgabe beziehen mehr als die Hälfte aller Schülerinnen und Schüler alle zur Verfügung stehenden Kriterien bei

1 Der Levene-Test wird mit $df1 = 3$ und $df2 = 846$ und $F_I = 18{,}9$ und $F_{II} = 10{,}9$ signifikant ($p<.001$), weshalb von inhomogenen Varianzen auszugehen ist.

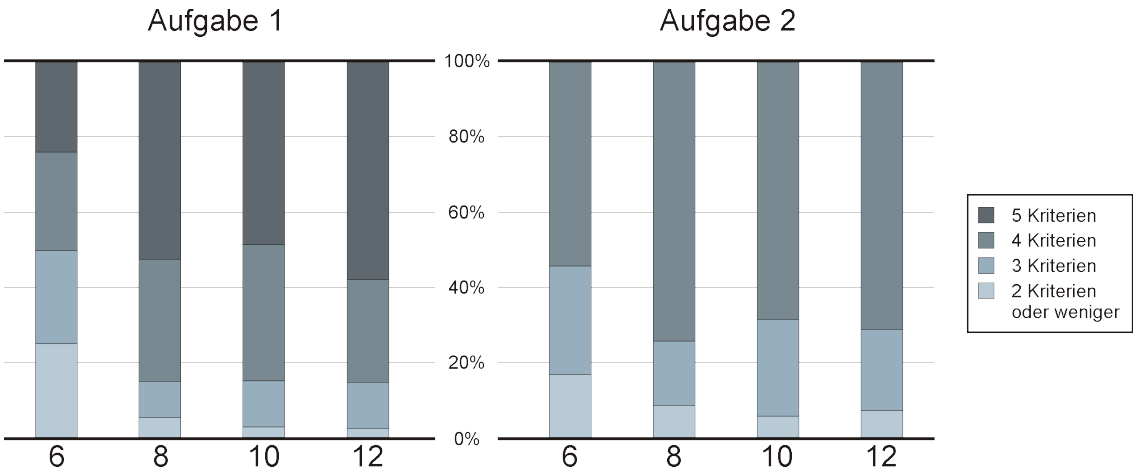

Abbildung 6.4: Anzahl der von Schülerinnen und Schüler verwendeten Kriterien in den Entscheidungsaufgaben nach Jahrgängen. Links: Aufgabe 1 (Energieerzeugung). Rechts: Aufgabe 2 (Energiespeicherung).

der Diskussion mindestens einer Option in ihre Bewertung mit ein.

6.3.2 Gewichten von Kriterien

Aus den erhobenen Daten lässt sich ebenfalls feststellen, dass der Anteil der Schülerinnen und Schüler, die das Vornehmen einer Gewichtung in ihrem Bewertungsprozess beschreiben, mit zunehmendem Schulalter ansteigt: Es bestehen hier geringe Korrelationen von $r = 0{,}33^{***}$ in der ersten und $r = 0{,}23^{***}$ in der zweiten Entscheidungsaufgabe.

Auch hier lässt sich ergänzend noch eine Analyse mittels geplanter Kontraste durchführen, die zu folgendem Ergebnis kommt: Alle geplanten Kontraste[1] werden hinsichtlich der Verwendung von Gewichtungen für die Kriterien bis auf eine Ausnahme signifikant: In der zweiten Aufgabe kann hinsichtlich des Gewichtens von Kriterien nicht mehr eindeutig zwischen Schülerinnen und Schülern der 10. Klasse und der Oberstufe unterschieden werden (siehe Tabelle 6.9).

Zusammenfassend kann jedoch festgehalten werden, dass Schülerinnen und Schüler mit zunehmendem Schulalter auch zunehmend die Fähigkeit besitzen, einzelne Kriterien in ihrem Bewertungsprozess zu gewichten und dies auch in ihrer Antwort zu dokumentieren. Insgesamt enthalten 24% der Antworten zu Aufgabe 1 und 18% der Antworten zu Aufgabe 2 Überlegungen zum Gewichten von Kriterien innerhalb des Bewertungsprozesses. Die

1 Der Levene-Test wird sowohl für die gesamte Stichprobe signifikant ($p<.001$) als auch für die paarweise untersuchten Gruppenkontraste (Klasse 6 zu Klasse 8, Klasse 8 zu Klasse 10 und Klasse 10 zu Oberstufe) signifikant mit $p<.01$, weshalb in diesem Fall von inhomogenen Varianzen auszugehen ist.

Tabelle 6.9: Kontraste hinsichtlich des Gewichtens von Kriterien.

Kontrast	*Gewichten bei Aufgabe 1*			*Gewichten bei Aufgabe 2*		
	Kontrastwert[a]	*Fehler*[b]	*Signifikanz*	*Kontrastwert*[a]	*Fehler*[b]	*Signifikanz*
Klassen 6 zu 8	+0,09	0,03	** p=.003	+0,09	0,03	*** p<.001
Klasse 8 zu 10	+0,15	0,04	*** p<.001	+0,08	0,04	* p=.028
Klasse 10 zu Oberstufe	+0,15	0,05	** p=.002	+0,06	0,04	p=.145

[a] Differenz der Mittelwerte in den jeweiligen Gruppen
[b] Standardfehler

häufigsten Gewichtungen erfolgen durch Aufwerten/Hervorheben eines oder mehrerer Kriterien, so dass sie andere Kriterien dominieren (ca. 65% aller Gewichtungen erfolgen auf diese Weise). Aber auch die umgekehrte Richtung, das Abwerten/Herunterstufen der Wichtigkeit von Kriterien (ca. 25% aller Gewichtungen erfolgen auf diese Weise) sowie Kombinationen (Aufwerten des einen und Abwerten eines anderen Kriteriums, ca. 10% aller Gewichtungen erfolgen auf diese Weise) werden des Öfteren beobachtet. In Einzelfällen (zweimal in der ersten und dreimal in der zweiten Aufgabe) überlegen die Schülerinnen und Schüler, ob eine Gewichtung von Kriterien angebracht ist oder nicht und entscheiden sich dann explizit für eine Gleichgewichtung aller Kriterien. Auch dieses Verhalten wurde als „Gewichtung“ codiert, weil diese Schülerinnen und Schüler die entsprechenden Überlegungen vornehmen.

KAPITEL 7

Diskussion

7.1 Reliabilität und Validierung

Reliabilitäten

Das entwickelte Messinstrument weist Reliabilitäten von 0,65 (Cronbach-Alpha) und 0,64 (WLE) auf. Sie sind damit vergleichbar mit aus anderen anderen Studien bekannten Reliabilitäten physikalischer Leistungstests, z. B. 0,65 (EAP/PV) beim „Kompetenzentwicklungstest“ von Viering (2012: S. 101) und 0,71 (Cronbach-Alpha) bei einem physikalischen Leistungstest von Kauertz (2007: S. 93ff), dessen Itempool aus veröffentlichten TIMSS-, PISA- und eigenen Aufgaben mit Bezug zum nordrhein-westfälischen Lehrplan für die Sekundarstufe I besteht (Kauertz 2007: S. 75f). Die WLE-Reliabilität des Naturwissenschaftstests von PISA2003 erreicht 0,77 (Prenzel et al. 2006: S. 98). Vorausgehende Studien zum *Bewerten, Entscheiden und Reflektieren* im Kontext nachhaltiger Entwicklung berichten bei Einsatz eines analogen Testinstruments in der Biologie eine EAP/PV-Reliabilität von 0,70 (Eggert & Bögeholz 2010: S. 244). Die von Hostenbach (2011: S. 98) vorgestellten Tests zur Messung von Bewertungskompetenz in den Unterrichtsfächern Biologie und Chemie weisen eine EAP/PV-Reliabilität von 0,65 bzw. 0,68 auf. Eine Erhöhung der in diesem Testinstrument erzielten Reliabilität kann bereits aus einer höheren Anzahl von Items resultieren, wie sie am Ende des nächsten Kapitels vorgeschlagen wird.

Schulalter

Bewertungskompetenz im Sinne des hier verwendeten Testinstruments im Kontext nachhaltiger Entwicklung steht in einem direkten Zusammenhang mit dem Schulalter, was einen Rückschluss auf die Kriteriumsvalidität erlaubt: Schülerinnen und Schüler höherer Jahrgangsstufen erreichen signifikant höhere Personenfähigkeiten (WLE-Schätzer) als die jeweils jüngeren Altersstufen. Dies ist im Einklang mit Befunden aus früheren Studien, bei denen sich allerdings nur die Kontraste von Klasse 6 zu Klasse 8 und von Klasse 10 zu Klasse 12 als signifikant herausgestellt haben (Eggert 2008, Eggert & Bögeholz

2010: S. 247 bzw. S. 108). In diesen Studien wurde als Ergebnis einer linearen Regressionsanalyse $r^2 = 0{,}28^{***}$ gefunden, im Rahmen dieser Studie kann dieser Befund für *Bewerten, Entscheiden und Reflektieren* im Physikunterricht mit $r^2 = 0{,}26^{***}$ nahezu reproduziert werden.

Schulnoten

Die mit $r = 0{,}27^{***}$ bzw. $r = 0{,}26^{***}$ höchsten Korrelationen der erhobenen Schulfächer in der 10. Klasse für Mathematik und Deutsch (siehe Tabelle 6.5) lassen sich vermutlich damit erklären, dass es in diesen Fächern vermehrt auf argumentative Kompetenzen und das Abwägen von Vor- und Nachteilen ankommt (KMK 2004), die auch bei der Bearbeitung von Aufgaben im Kontext nachhaltiger Entwicklung hilfreich sein können. Zusammengenommen weist die Deutschnote in der Sekundarstufe I den größten Zusammenhang mit *Bewerten, Entscheiden und Reflektieren* auf (siehe Tabelle 6.5, rechts). Auch die Mathematiknote kann möglicherweise mit argumentativen und Logik-Kompetenzen erklärt werden. Nicht zu erklären ist allerdings, warum diese Korrelationen bei den Schülerinnen und Schülern der Oberstufe wieder geringer ausfallen. Stattdessen sind in der Oberstufe nur signifikante Korrelationen zu den beiden Naturwissenschaften erkennbar, wobei die Korrelation zur Physiknote am größten ausfällt. Bei nur 95 Fällen ist allerdings auch klar, dass es sich aufgrund der beschriebenen Kurswahleffekte in der Oberstufe hier vornehmlich um technisch besonders interessierte Schülerinnen und Schüler handeln könnte.

Die im Vergleich zur Biologienote geringere Korrelation der Physiknote in den Klassen 6 bis 10 kann auch ein Hinweis darauf sein, dass Bewerten im Physikunterricht vernehmlich (noch) als innerfachliches Bewerten (Schecker & Höttecke 2007) verstanden wird und sich somit weniger in der Physiknote widerspiegelt als in der Biologienote. Diese beinhaltet auch (schon) vermehrt Komponenten des gesellschaftlichen Lebens und der Nachhaltigkeit (siehe Tabelle 2.1). In den Studien von Eggert (2008), Eggert und Bögeholz (2010) wurden ebenfalls Korrelationen zu Schulnoten berechnet[1]. Signifikante Korrelationen ergeben sich hier ebenfalls für die Zusammenhänge von *Bewerten, Entscheiden und Reflektieren* mit der Deutsch- ($r = 0{,}23^{*}$) und Mathematiknote ($r = 0{,}24^{*}$), ein Zusammenhang mit der Politik- oder Physiknote wird in diesen Studien nicht berichtet.

Schulnoten bilden immer mehrere Kompetenzen von Schülerinnen und Schülern in den jeweiligen Unterrichtsfächern ab. Fasst man sie als einen Indikator für Fachwissen auf, so lässt sich beim Blick auf die rechte Spalte in Tabelle 6.5 insgesamt feststellen und hervorheben, dass aufgrund der nur geringen Korrelationen kein relevantes physikalisches Fachwissen für die Bearbeitung des Testinstruments nötig ist, was auch für die Qualität der Aufgabenpräsentation spricht. Für die Unterrichtsfächer Biologie und Chemie werden

1 Allerdings für alle untersuchten Jahrgänge (also auch der Oberstufe) – wiederum einzeln z-standardisiert.

in Hostenbach (2011: S. 114) Korrelationen zwischen den Dimensionen *Kenntnis von Fachwissen* und *Bewertungskompetenz* von $r = 0{,}15^{***}$ bzw. $r = 0{,}14^{***}$ berichtet und zwischen den Dimensionen *Umgang mit Fachwissen* und *Bewertungskompetenz* $r = 0{,}22^{***}$ bzw. $r = 0{,}30^{***}$. Diese Korrelationen liegen damit in ähnlichen Größenordnungen, wie sie bereits durch Schulnoten abgebildet werden.

Lesekompetenz

Ein nachweislicher Zusammenhang von Lesegeschwindigkeit mit *Bewerten, Entscheiden und Reflektieren* besteht nicht (siehe Tabelle 6.6). Aus den Erfahrungen der Erhebung ist bekannt, dass nur in Einzelfällen Schülerinnen und Schüler aus Zeitmangel nicht mehr alle Aufgaben im hinteren Teil des Testhefts bearbeiten konnten (siehe Kapitel 5.2). Ein möglichst geringer Zusammenhang zwischen Leseverständnis und Testergebnis für *Bewerten, Entscheiden und Reflektieren* spricht für eine bessere Abgrenzung gegenüber Lesekompetenz. Für das im Rahmen dieser Arbeit entwickelte Testheft konnte eine nur sehr schwache (partielle) Korrelation zwischen Leseverständnis und *Bewerten, Entscheiden und Reflektieren* von $r = 0{,}12^{***}$ festgestellt werden ($N = 850$), der Einfluss von Leseverständnis auf *Bewerten, Entscheiden und Reflektieren* ist damit sehr schwach ($r^2 = 0{,}01$). Diese Korrelationen liegen damit deutlich unterhalb derer aus anderen Studien: So berichtet Viering (2012: S. 136) Korrelationen zwischen seinen Testergebnissen und Lesegeschwindigkeit bzw. Leseverständnis von $r = 0{,}09^{***}$ bzw. $r = 0{,}20^{***}$. In gezielten Untersuchungen zur Rolle von Lesegeschwindigkeit und Leseverständnis bei naturwissenschaftlichen Kompetenztests berichtet Hartmann (2013: S. 105) deutlich höhere Korrelationen zwischen einem *Kompetenztest Biologie* und Lesegeschwindigkeit bzw. Leseverständnis von $r = 0{,}21^{*}$ bzw. $r = 0{,}44^{***}$. Vor dem Hintergrund dieser Zusammenhänge in anderen Studien kann daher der in dieser Studie verwendete Ansatz bei der Aufgabenkonstruktion, den Einfluss von Lesekompetenz auf das Testergebnis durch Wahl einer tabellarischen Aufgabenrepräsentation so weit wie möglich zu minimieren (siehe Kapitel 4.2), als gelungen bezeichnet werden.

Schlussfolgerungen

mögliche Forschungsansätze

Es fehlt derzeit an Studien, die gezielt Verbindungen zwischen physikalischem Fachwissen und Bewertungskompetenz in Physik herstellen. Hierzu bedarf es weiterer Arbeiten, die diese Dimensionen gemeinsam erheben und untersuchen.

Weitere Hinweise auf Konstruktvalidierung können erzielt werden, wenn in weiteren Studien das hier entwickelte Testinstrument im Sinne einer Methodentriangulation um weitere Methoden wie z. B. Interviews ergänzt wird (vgl. Gläser-Zikuda et al. 2012). Ebenso könnten Bewertungskompetenztests aus anderen Unterrichtsfächern eingesetzt werden, um domänenspezifische Gemeinsamkeiten und Unterschiede heraus zu arbeiten.

7.2 Modellpassung und Item-Fit-Parameter

In dieser Arbeit wurden alle erhobenen Daten auch ausgewertet und in die Modellierung einbezogen (mit Ausnahme der in Kapitel 5 genannten 7 Schülerinnen und Schüler, die das Testheft nicht ernsthaft ausgefüllt bzw. das Ausfüllen verweigert haben), ohne Schülerinnen und Schüler aufgrund zu schlecht passender Personenparameter im Nachhinein von der Studie auszuschließen.

eindimensionales Raschmodell

Ausgehend von der Zielsetzung der Anwendung des eindimensionalen Rasch-Partial-Credit-Modells, welches in früheren Studien bereits erfolgreich für die Modellierung von Bewertungskompetenz im Kontext nachhaltiger Entwicklung und die Teilkompetenz *Bewerten, Entscheiden und Reflektieren* angewendet werden konnte (Eggert 2008, Eggert & Bögeholz 2010), konnten Zusammenlegungen von Kategorien gefunden werden, die zusammengenommen sehr zufriedenstellende Item-Fit-Parameter für das Rasch-Partial-Credit-Modell aufweisen. Dieses Modell kann daher die empirisch gewonnenen Daten gut erklären und repräsentieren. Die gute Passung weist zudem darauf hin, dass *Bewerten, Entscheiden und Reflektieren* im Rahmen des hier entwickelten Testinstruments als eindimensionales Konstrukt erfasst wird.

Hingegen kann nicht ausgeschlossen werden, dass die Zusammenlegungen der Kategorien die optimale Lösung dahingehend sind, als dass andere Modelle (wie z. B. Mixed-Rasch-, latente-Klassen- oder Hybrid-Modelle) mit anderen Zusammenlegungen zu einer noch besseren Repräsentation der Daten kommen. Studien, die Vorarbeiten für die ESNaS-Bewertungskompetenz darstellen, modellieren diese ebenfalls eindimensional. Im Gesamtmodell ordnen sie dabei allerdings den beiden untersuchten Unterrichtsfächern *Biologie* und *Chemie*, sowie dem Konstrukt *außerfachliche Bewertung* jeweils eine eigene Dimension zu (Hostenbach 2011: S. 97).

Bei einer Untersuchung des Zusammenhangs zwischen *Bewertungskompetenz* und *Problemlösen* im Fach Chemie weist Heitmann (2012: S. 141) diesen Kompetenzen ebenfalls jeweils eine eigene Dimension zu: Modellparameter zeigen hier, dass ein zweidimensionales Rasch-Partial-Credit-Modell auch besser auf die Daten passt als ein eindimensionales und damit, dass Bewertungskompetenz und Problemlösekompetenz über vergleichsweise eigenständige Kompetenzanteile verfügen. Allerdings lassen sich Korrelationen von r=0,76*** zwischen ihnen feststellen.

verschiedene Auswerteprogramme

In dem Bestreben, die erhaltenen Daten von möglichst vielen Facetten zu beleuchten und durch diesen umfassenden Blick auch eine große Aussagekraft zu erhalten, wurden zwei verschiedene Programme zur Raschmodellierung eingesetzt: ConQuest und WINMIRA.

Hauptunterschied zwischen diesen ist, dass WINMIRA nur mit kompletten Datensätzen arbeitet (also einen listenweisen Ausschluss vornimmt, siehe Kapitel 5.2) während ConQuest individuell für jede Aufgabe alle verfügbaren Daten einbezieht (fallweiser Ausschluss, siehe Kapitel 5.2). Diese beiden Programme arbeiten jedoch noch zusätzlich mit unterschiedlichen Algorithmen, so dass z. B. die Berechnung von Thresholds bei WINMIRA zu anderen Ergebnissen kommt als die Item-Deltas von ConQuest – selbst wenn man beiden Programmen identische und vollständige Datenmatrizen zur Verfügung stellt. In Bezug auf den intendierten umfassenden Blick auf die empirischen Daten müssen aber leider Abstriche gemacht werden, da beide Programme im Hinblick auf manche Items leider nicht zu einheitlichen Aussagen kommen. Allerdings kann aus diesem Aspekt auch der Vorteil gezogen werden, dass keine der im Test verbleibenden Items sich in beiden Programmen durchgängig so auffällig zeigt, als dass es unbedingt aus dem Test entfernt werden sollte.

Diskussion einzelner Items

Die Items 3 und 4 (Optionen „Nordsee Offshore“ und „Ostsee Offshore“) sind sich eventuell noch zu ähnlich und könnten in weiteren Studien dahingehend abgeändert werden, dass sie sich stärker unterscheiden (evtl. genügt schon eine einfache Streichung des „Offshore“-Beisatzes in der Bezeichnung der Option). Der T-Wert indiziert hier, dass der WMNSQ-Index zwar nicht mehr innerhalb des jeweils berechneten 95%-Konfidenzintervalls liegt, die WMNSQ-Werte bleiben allerdings deutlich innerhalb der empfohlenen Grenzwerte. Bond und Fox (2010: S. 240) heben zudem hervor, dass Overfits nicht so problematisch wie Underfits sind. Es kommt hinzu, dass innerhalb einer Entscheidungsaufgabe nicht einfach eine einzelne Option unbeachtet bleiben kann, bei zu schlechten Item-Fit-Parametern müsste die gesamte Aufgabe aus der Berechnung der Gesamtskala herausgehalten werden.

Die Beschreibung der dargestellten Strategien mittels MC-Aufgaben weist die erwartete Zunahme der Aufgabenschwierigkeit von intuitiv-rechtfertigend über non-kompensatorisch bis kompensatorisch auf. Nicht zuletzt ist Item 13 (Beschreibung der kompensatorischen Strategie) unter diesen deswegen das schwierigste, weil es zum Auffinden des Attraktors unter den z. T. recht ähnlich formulierten Distraktoren von Nöten ist, das Vorgehen der Person genau verstanden zu haben.

kohärentes Scoring

Das in Item 16 angewandte Scoring (siehe Tabelle 5.2) resultiert bei Itemstep 16.1 in einer verglichen mit Item 15.1 leichteren Item-Schwierigkeit. In weiteren Auswertungen sollte überprüft werden, ob es unabhängig von besseren Item-Fit-Parametern nicht doch zweckmäßiger ist, die Kategorie „nichts verbessern“ (siehe Kapitel 5.1.2) mit dem Score 0 zu versehen (siehe Hinweis in Kapitel A.3).

Schlüsse aus den multiplen Imputationen

Auffälligkeiten von Items zur kompensatorischen Strategie verfestigen sich bei den mittels multipler Imputation vervollständigten Datensätzen dahingehend, dass zunächst nur Item 13 (Beschreibung der kompensatorischen Strategie) einen Underfit beim Z_Q-Index aufweist, der sich jedoch nicht im T-Wert des WMNSQ widerspiegelt (siehe Tabelle 6.7). Bei den mittels multipler Imputation vervollständigten Datensätze werden diese T-Werte in drei der beispielhaft durchgeführten Imputationen auffällig – allerdings widersprechen die zwei anderen Imputationen einem einheitlichen Bild (siehe Tabellen A.1 bis A.5).

Bei den vervollständigten Datensätzen kann zudem beobachtet werden, dass die Items 15 (Verbesserungsvorschlag für die non-kompensatorische Strategie) und 16 (Verbesserungsvorschlag für die kompensatorische Strategie) beim Z_Q-Index nun ebenfalls auffällig werden und hier einen Underfit aufweisen, die sich allerdings ebenfalls nicht im T-Wert des WMNSQ wiederholen (siehe Tabellen A.1 bis A.5). Item 16 weist zusätzlich zwei weitere Auffälligkeiten im Original-Datensatz auf, genau wie das Item 17: Die Item-Thresholds (ConQuest) sind beim Score 2 tendenziell etwas zu schwierig, zudem sind die Pt Bis für den Score 2 zwar nicht mehr geordnet, aber weiterhin positiv (siehe Tabelle 6.7). Dies liegt möglicherweise darin begründet, dass diese Kategorien auch nur selten vergeben wurden ($n_{16.2}$=36 und $n_{17.2} = 11$, siehe Tabelle 6.1: ursprüngliche Scores 4 bzw. 3). Der leicht zu hohe Item-Threshold (ConQuest) bei Item 16.2 verschwindet bei einer der mittels multipler Imputation vervollständigten Datensätze wieder und fällt hier kleiner als 3,00 aus (siehe Tabelle A.1), der ebenfalls geringfügig zu hohe Item-Threshold (ConQuest) des Items 17.2 bleibt jedoch bestehen (siehe Tabellen A.1 bis A.5). Im Falle der Nicht-Ordnung der Pt Bis bei den Scores 2 für Items 16 und 17 ergeben sich keine Veränderungen und Tendenzen aus den mittels multipler Imputation vervollständigten Datensätzen (siehe Tabellen A.1 bis A.5).

Gesamtwürdigung

In der Gesamtwürdigung bleibt also festzustellen, dass die Item-Fit-Parameter der Entscheidungsaufgaben weitestgehend unauffällig sind: 8 Items eignen sich hinsichtlich aller aufgestellten Kriterien (siehe Kapitel 5.3.2) sehr gut zur Verwendung in einem eindimensionalen Rasch-Partial-Credit-Modell. Kein Item fällt hinsichtlich der Item-Fit-Parameter durchgängig schlecht auf. Allerdings gibt es bei den anderen 8 Items kleinere Einschränkungen, die zwar genannt und diskutiert werden, aber in der durchgeführten Gesamtabwägung mit dem Mehrgewinn an Informationen für das Modell nicht so gravierend erscheinen, als dass sie aus dem Testinstrument entfernt werden sollten. Vielmehr ist das Beibehalten dieser Items sehr wichtig für ein konsistentes Gesamtbild, das alle Optionen der Entscheidungsaufgaben und die Beschreibungen und Verbesserungsvorschläge aller vorgestellten Strategien einbezieht.

Item-Fit-Parameter bei anderen Studien

Im Vergleich zu anderen Studien aus der empirischen Kompetenzforschung, die ebenfalls Raschmodellierungen vornehmen, liegt eine Besonderheit dieser Studie darin, die erhaltenen Ergebnisse zum Testinstrument mit zwei Auswerteprogrammen über verschiedene Verfahren abzusichern. Vielfach wird im Zuge der Raschmodellierung nur der WMNSQ- und der dazugehörige T-Wert betrachtet, der Bedarf an zusätzlichen Fit-Parametern wird nach Bond (2005: S. 336) primär im europäischen und weniger im amerikanischen Raum angemeldet. Auch die Studien von Viering (2012), Hostenbach (2011), Heitmann (2012) und Kauertz (2007) diskutieren primär WMNSQ-Werte (*Infit*). Desweiteren werden in manchen dieser Arbeiten – auch bedingt durch das jeweilige Testdesign – EAP/PV-Schätzer verwendet. Diese sind jedoch im Vergleich zu den in dieser Studie diskutierten WLE-Schätzern im Allgemeinen robuster gegen Ausreißer, können aber die Personenfähigkeit nicht so gut schätzen wie die WLE-Schätzer (vgl. Rost 2004: S. 314ff).

Schlussfolgerungen

mehrere Entscheidungsstrategien verwenden

Im Rahmen der Konstruktion dieses Testinstruments wurden als Repräsentant nonkompensatorischer Strategien die *Elimination by Aspects*-Regel gewählt, als Repräsentant kompensatorischer Strategien die *Weighted Additive Rule* (siehe Tabelle 1.1). Zur weiteren Differenzierung zwischen diesen beiden Klassen von Strategien könnte in weiteren Studien noch Aufgabenteile hinzu genommen werden, die auf der *Lexicographic Rule* (siehe Tabelle 1.1, Fishburn 1974, Goldstein & Gigerenzer 2002) basieren, da sie optimal an das bereits verwendete Muster mit Schilderung von Wichtigkeiten und dargebotenen Entscheidungen anschließt. Um die Bearbeitungszeit in Grenzen zu halten, sollten dann allerdings Überlegungen zu unterschiedlich ausgestalteten Testheften angestellt werden.

Item-Anzahl vergrößern

Mit Blick auf die Verteilung von Items in der Person-Item-Map kann der Vorschlag gemacht werden, weitere Items in das Testheft einzubringen, die besonders den unteren Leistungsbereich stärker abdecken. Die Aufsplittung der MC-Aufgaben zur Beschreibung von Strategien in jeweils 2 MC-Aufgaben auf unterschiedlichem Niveau kann in dieser Hinsicht synergistisch wirken, da neben der erwarteten dichteren Itemverteilung im unteren Leistungsbereich dann auch stärkere Aussagen zum Verstehen und Nachvollziehen von Strategien durch Schülerinnen und Schüler gemacht werden können. Alternativ könnten die bestehenden MC-Aufgaben auch explizit als *Ordered MC-Aufgaben* (Briggs et al. 2006) mit mehreren Itemsteps konstruiert werden: Dabei spielt die Annahme mit ein, dass einzelne Antwortmöglichkeiten in einem bestimmten Fähigkeitsniveau vorwiegend ausgewählt werden. Im Rahmen der Entwicklung einer physikalischen *learning progression* zu Kraft und Bewegung konnten Ordered MC-Aufgaben von Alonzo und Steedle (2009) bereits gewinnbringend eingesetzt werden. Sie bilden damit zugleich auch eine Analogie zum

Scoring in den Verbesserungsvorschlag-Aufgaben, bei denen die Antworten der Schülerinnen und Schüler ebenfalls in Kategorien mit teilweise richtigen und vollständig richtigen Verbesserungsvorschlägen einsortiert werden.

7.3 Umgang mit Optionen und Kriterien

Gewichten anspruchsvoll

Zur Lage der Items der beiden Entscheidungsaufgaben in der Person-Item-Map (Abbildung 6.2) können zwei Feststellungen gemacht werden: Zum einen scheint die Aufgabe 1 (Energieerzeugung durch Windkraft) den Schülerinnen und Schülern als Kontext und auch bei der Diskussion der einzelnen Optionen deutlich vertrauter zu sein als Aufgabe 2 (Energiespeicherung), da die dazugehörigen Items 1 bis 4 in der Aufgabe 1 im Mittel leichter sind als die gemittelte Aufgabenschwierigkeit der Items 6 bis 9 in der Aufgabe 2. Zum anderen ist das Gewichten von Kriterien anspruchsvoller als die Diskussion der Optionen: Die Items 5 und 10 sind jeweils die schwierigsten in den beiden Entscheidungsaufgaben, zudem ist das Gewichten von Kriterien in Aufgabe 2 (Item 10) schwieriger als in Aufgabe (Item 5). Dass die Gewichtung von Kriterien schwieriger ist als die Diskussion der einzelnen Optionen ist im Einklang mit den Ergebnissen anderer Studien (Eggert & Bögeholz 2010, Jimenez-Aleixandre 2002, Kolstø 2006, Seethaler & Linn 2004). Hinsichtlich des Gewichtens von Kriterien in Entscheidungsprozessen muss jedoch darauf hingewiesen werden, dass nur explizite Gewichtungen der Schülerinnen und Schüler als solche erkannt und mit dem entsprechenden Score versehen wurden. Die wahre Anzahl an vorgenommener Gewichtungen wird somit im Rahmen dieser Studie wohl unterschätzt, weil implizite Gewichtungen mit den hier eingesetzten Methoden nicht erfasst werden können.

Vertrautheit der Kontexte

Der Kontext der ersten Entscheidungsaufgabe (Erzeugung elektrischer Energie) ist den Schülerinnen und Schülern aus ihren Alltagserfahrungen heraus sicherlich viel präsenter als der Kontext der zweiten Entscheidungsaufgabe (Energiespeicherung). Als einer von acht Kontexten der Konzeptionalisierung *Umwandlung und Transport* schneidet die Windenergie in einer Studie zur Entwicklung eines physikalischen Kompetenzentwicklungstests von Viering (2012: S. 105) ebenfalls als zweitleichtester hinsichtlich der Aufgabenschwierigkeit ab.

Kriterien fachwissenschaftlich belegt

Bei der Recherche zu den dargebotenen Optionen und der Darstellung der Kriterien wurde darauf Wert gelegt, diese mit Erkenntnissen aus der Fachwissenschaft zu untermauern. Die beschriebenen Optionen können allerdings aufgrund ihrer Komplexität in diesem Testheft in ihren Kriterien nicht abschließend und umfassend beschrieben werden und müssen für den Schuleinsatz didaktisch aufbereitet werden. Dies geschah während der Aufgabenentwicklung auch unter dem Aspekt, gemeinsame Kriterien für alle Optionen zu finden. Aus technischer

Sicht kommen daher zu jeder Option ggf. noch Spezifika hinzu, die hier keine Erwähnung finden konnten: Beispielhaft sei die potentielle Nutzung von Wasserstoff als Kraftstoff in Automobilen angesprochen, die diese Option natürlich im Hinblick auf Mobilität interessanter macht. Ebenso mussten das Kriterium *Sicherheitsrisiken* in der zweiten Entscheidungsaufgabe auf ein für Schülerinnen und Schüler verständliches Maß qualitativ beschränkt werden.

Häufigkeiten verwendeter Kriterien

In einer Untersuchung zum Zusammenhang von *Bewertungskompetenz* und *analytischem Problemlösen* bei Schülerinnen und Schülern der 10. Klasse (Gymnasium, N=612 bzw. N=741) hat auch Heitmann (2012) eine Kategorisierung von Kriterien in ihren Testmaterialien vorgenommen. *Naturwissenschaftliche* Kriterien werden hier in 78% aller Bewertungen verwendet (Heitmann 2012: S. 148), was geringfügig unterhalb der in dieser Studie gefundenen Ergebnisse für Schülerinnen und Schüler der 10. Klasse (Gymnasium, N=222) liegt: Die Kriterien *Energieverlust beim Transport* in der ersten und *Speichergüte* in der zweiten Entscheidungsaufgabe können ebenfalls einer solchen Kategorie zugerechnet werden und werden in 83% bzw. 91% aller Entscheidungen verwendet (siehe Tabelle 6.8). Bei anderen Kriterien, wie z. B. denen der Kategorien *nachhaltige Entwicklung* oder *subjektiv-emotional*, werden in der Studie von Heitmann (2012: S. 148) durchweg viel geringere Häufigkeiten berichtet: Nur 43% aller Schülerinnen und Schüler beziehen sich auf Kriterien der Kategorie *nachhaltigen Entwicklung*. In der hier vorliegenden Studie lassen sich am ehesten die Kriterien *Naturschutz* und *Sicherheitsrisiken* zu dieser Kategorisierung von Heitmann (2012) zählen und werden von Schülerinnen und Schülern der 10. Klasse in 86% bzw. 88% aller Bewertungen verwendet (siehe Tabelle 6.8). 40% der Schülerinnen und Schüler in der Studie von Heitmann (2012) beziehen sich auf *subjektiv-emotionale Kriterien*. In der vorliegenden Studie lässt sich das Kriterium *Sichtbarkeit/Lärm* zu dieser Kategorie zählen und wird von Schülerinnen und Schülern der 10. Klasse in 81% aller Bewertungen verwendet.

Auch ein Vergleich mit weiteren Studien zur Verwendung von Kriterien in Bewertungsprozessen von Studierenden zeigt, dass Kriterien zur nachhaltigen Entwicklung recht selten verwendet werden: Uskola et al. (2010: S. 2320) berichten bei Beobachtungen von Diskussionen von $N = 15$ Studierenden über die Wahl eines Heizsystems, dass bei einer Gesamtanzahl von 802 Kriterien-Nennungen zwar 237 Nennungen der Kategorie *ökonomisch* zugeordnet werden können, allerdings nur 36 Nennungen der Kategorie *Nachhaltigkeit*. Halverson et al. (2009) berichten über die Auswertung von Aufsätzen zur Stammzellenforschung, die $N = 132$ Studierende der Biologie nach selbstverantworteten Internet-Recherchen verfasst haben: Kriterien der Kategorie *Medical* (die beispielsweise Vorteile medizinischer Behandlungen aber auch Gesundheitsrisiken umfasst) werden in 74%, ökonomische Kriterien in 23% aller Fälle mit einbezogen. Andere Kategorien von

Kriterien aus diesen Studien lassen sich nicht für einen Vergleich der verwendeten Kriterien in dieser Studie heranziehen. Dass in dieser Studie durchweg höhere Verwendungen von Kriterien zu verzeichnen sind, ist sicherlich auch durch das hier verwendete Studiendesign bedingt: Mögliche Kriterien werden den Schülerinnen und Schülern im Informationsheft bereits aufbereitet in Tabellenform vorgelegt und müssen nicht erst aus vorliegendem Material generiert werden.

Schlussfolgerungen

Rolle impliziter Gewichtungen

Die Frage, ob Schülerinnen und Schüler neben den in ihren Antworten geäußerten auch noch implizite Gewichtungen vornehmen, könnte begleitend durch z. B. speziell konstruierte Aufgaben in weiteren Studien gemessen werden. Dabei werden diese Aufgaben so gestaltet, dass die Wahl der Option einen Rückschluss darauf ermöglicht, ob die Schülerinnen und Schüler eine Gewichtung vorgenommen haben. Alternativ könnte man Schülerinnen und Schüler direkt z. B. mittels Ratingskalen oder im Anschluss durchgeführter Kurz-Interviews danach befragen, ob sie eine Gewichtung in ihren Entscheidungen vorgenommen haben (vgl. Jungermann et al. 2005: S. 137, 414ff). Die hieraus gewonnen Ergebnisse tragen wiederum zu einem tieferem Verständnis von Gewichtungen in Entscheidungen von Schülerinnen und Schülern bei.

7.4 Generalisierbarkeit der Studie

Stichprobenwahl

Die Studie wurde in Niedersachsen an drei Gymnasien durchgeführt, davon eins im ländlichen Bereich, eins im Ballungsraum einer Großstadt und eins in einer Großstadt. Streng genommen können daher im Rahmen dieser Studie nur Aussagen zum *Bewerten, Entscheiden und Reflektieren* von Schülerinnen und Schülern dieser drei Schulen gemacht werden. Die Auswahl der Schulen erfolgte nicht repräsentativ, die Auswahl der getesteten Klassen quasi-repräsentativ durch externe, z. T. zufallsbedingte Faktoren (z. B. in Klassen, in denen ansonsten Unterricht durch eine fehlende Fachlehrkraft ausgefallen wäre). Es kann daher auch mit Verweis auf die im Verhältnis zu anderen Studien recht große Stichprobenmenge von $N = 850$ vermutet werden, dass sich bei weiteren Erhebungen die Kompetenz für *Bewerten, Entscheiden und Reflektieren* von Schülerinnen und Schülern auch anderer Gymnasien nicht bedeutsam von den hier berichteten Ergebnissen unterscheiden werden.

Die vorliegende Studie ist eine Querschnittuntersuchung und betrachtet die Teilkompetenz *Bewerten, Entscheiden und Reflektieren* im Kontext nachhaltiger Entwicklung in

physikalischen Kontexten zum Themenfeld Energiegewinnung, -speicherung und -nutzung in den Klassenstufen 6, 8, 10 und 12. Sie liefert damit erste Hinweise auf eine mögliche Entwicklung von *Bewerten, Entscheiden und Reflektieren* während der Schulzeit. Die wahren Entwicklungsverläufe können allerdings nur in Längsschnittstudien betrachtet werden, wenn dieselben Schülerinnen und Schüler im Verlauf ihrer Schulzeit wiederholt getestet werden.

Aktualität

Die Haupterhebung fand im November 2010 statt. In den letzten drei Jahren hat das Thema *Energiewende* und *Erneuerbare Energien* eine große mediale Präsenz erfahren und wird auch außerhalb von Schule und Forschung mit Blick auf die Gestaltung unserer Zukunft breit diskutiert. Neben der Eröffnung von Offshore-Windenergieparks (z. B. alpha ventus 2010) werden auch zunehmend verschiedene Speichermöglichkeiten diskutiert. Dazu zählen z. B. große Pumpspeicherseen in Norwegen (Asendorpf 2011) und Lithium-Akkus in Schwerin (WEMAG 2013).

Die gefundenen Kontexte sind daher sowohl gegenwarts- als auch zukunftsrelevant. Sie unterstreichen im Zusammenhang mit der Ausbildung von mündigen, reflektiert entscheidenden Bürgerinnen und Bürgern zugleich die elementare Wichtigkeit von Bewertungskompetenzen bei Schülerinnen und Schülern für zahlreiche Entscheidungen, die im Zusammenhang mit dem kontinuierlichen Ausbau Erneuerbarer Energiequellen gefällt werden müssen, und für eine Sensibilisierung zum nachhaltigen Handeln in ihrem eigenen Interesse und im Interesse ihrer Nachfahren.

Literaturverzeichnis

Abelson R.P., Levi A., 1985. Decision making and decision theory, Vol. 1. In: G. Lindzey; E. Aronson (Hrsg.), Handbook of social psychology, Seiten 231–309. Addison-Wesley, Reading, MA.

Alonzo A., Steedle J., 2009. Developing and Assessing a Force and Motion Learning Progression. Science Education, 93(3), 389–421, DOI: 10.1002/sce.20303.

alpha ventus, 2010. Deutschlands erster Offshore-Windpark alpha ventus wird feierlich eröffnet. Pressemitteilung vom 27.04.2010; online verfügbar unter http://www.alpha-ventus.de/uploads/media/alpha_ventus_PM_Eroeffnung_0427_de.pdf, abgerufen am 04.03.2013.

Andrich D., de Jong J.H.A.L., Sheridan B.E., 1997. Diagnostic Opportunities with the Rasch Model for Ordered Response Categories. In: J. Rost; R. Langeheine (Hrsg.), Applications of latent trait and latent class models in the social sciences, online verfügbar unter http://hbanaszak.mjr.uw.edu.pl/TempTxt/BOOK/c04.pdf, abgerufen am 15.08.2013, Seiten 59–72. Waxmann, New York, NY.

Asendorpf D., 2011. Norwegen, der Akku Europas. Die Zeit Ausgabe 36.

Baker F., 2001. The Basics of Item Response Theory. online verfügbar unter: http://info.worldbank.org/etools/docs/library/117765/Item Response Theory - F Baker.pdf, abgerufen am 19.02.2013.

Bakker R.H., Pedersen E., van den Berg G.P., Stewart R.E., Lok W., Bouma J., 2012. Impact of wind turbine sound on annoyance, self-reported sleep disturbance and psychological distress. Science of the Total Environment, 425, 42–51.

Baumert J., Klieme E., Neubrand M., Prenzel M., Schiefele U., Schneider W., Stanat P., Tillmann K.J., Weiß M., 2001. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich, Opladen.

Beck U., 2007. Weltrisikogesellschaft. Auf der Suche nach der verlorenen Sicherheit. Suhrkamp, Frankfurt am Main.

Betsch T., Funke J., Plessner H., 2011. Denken – Urteilen, Entscheiden, Problemlösen. Springer-Verlag, Berlin Heidelberg.

Betsch T., Haberstroh S., 2005. The Routines of Decision Making. Erlbaum Associates, Mahwah, NJ.

BGBl, 2002. Gesetz zur geordneten Beendigung der Kernenergienutzung zur gewerblichen Erzeugung von Elektrizität. Bundesgesetzblatt – Teil 1, Bundesanzeiger Verlag, (1341), Nr. 26, 1351–1359.

BGBl, 2010. Elftes Gesetz zur Änderung des Atomgesetzes. Bundesgesetzblatt – Teil 1, Bundesanzeiger Verlag, Nr. 62 vom 13.12.2010, 1814–1816.

BGBl, 2011. Dreizehntes Gesetz zur Änderung des Atomgesetzes. Bundesgesetzblatt – Teil 1, Bundesanzeiger Verlag, Nr. 43 vom 06.08.2011, 1704–1705.

Bilgili M., Yasar A., Simsek E., 2011. Offshore wind power development in Europe and its comparison with onshore counterpart. Renewable and Sustainable Energy Reviews, 15, 905–915.

Bodzin A., 2012. Investigating Urban Eighth-Grade Students' Knowledge of Energy Resources. International Journal of Science Education, 34(8), 1255–1275.

Bögeholz S., 2000. Natur erleben und gestalten. Politische Ökologie, Sonderheft 12, 17–18.

Bögeholz S., 2006. Explizites Bewerten und Urteilen. Beispielkontext Streuobstwiese. Praxis der Naturwissenschaften – Biologie in der Schule, 55(1), 89–115.

Bögeholz S., 2007. Bewertungskompetenz für systematisches Entscheiden in komplexen Gestaltungssituationen Nachhaltiger Entwicklung. In: D. Krüger; H. Voigt (Hrsg.), Theorien in der biologiedidaktischen Forschung, Seiten 209–220. Berlin, Springer.

Bögeholz S., 2011. Bewertungskompetenz im Kontext Nachhaltiger Entwicklung: Ein Forschungsprogramm. In: D. Höttecke (Hrsg.), Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie, Seiten 32–46. LIT-Verlag, Berlin.

Bögeholz S., Barkmann J., 2005. Rational Choice and beyond: Handlungsorientierende Kompetenzen für den Umgang mit faktischer und ethischer Komplexität. In: R. Klee; A. Sandmann; H. Vogt (Hrsg.), Lehr- und Lernforschung in der Biologiedidaktik, Band 2, Seiten 211–224. Studienverlag, Innsbruck.

Bonß W., 1996. Die Rückkehr der Unsicherheit: Zur gesellschaftstheoretischen Bedeutung des Risikobegriffs. In: G. Banse (Hrsg.), Risikoforschung zwischen Disziplinarität und Interdisziplinarität. Von der Illusion der Sicherheit zum Umgang mit Unsicherheit., Seiten 165–184. Berlin.

Bond T., 2005. Past, present and future: An idiosyncratic view of Rasch measurement. In: S. Alagumalai; D. Curtis; N. Hungi (Hrsg.), Applied Rasch Measurement: A book of exemplars, Seiten 329–341. Springer, Dordrecht.

Bond T., Fox C., 2010. Applying the Rasch Model. Lawrence Erlbaum Associates Inc, Mahwah, NJ, and Routledge (Reprint), NY.

Bortz J., Döring N., 2002. Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 3. Aufl., Springer-Verlag, Berlin Heidelberg New York.

Brügelmann H., 2001. Kontroversen um die Schulleistungsmessung in Deutschland. In: F.E. Weinert (Hrsg.), Leistungsmessung in Schulen, Seiten 33–44. Beltz Verlag, Weinheim und Basel.

Briggs D., Alonzo A., Schwab C., Wilson M., 2006. Diagnostic Assessment With Ordered Multiple-Choice Items. Educational Assessment, 11 (1), 33–63.

Bühner M., 2006. Einführung in die Test- und Fragebogenkonstruktion. Pearson, München.

Bühner M., 2012. Einführung in die Test- und Fragebogenkonstruktion, 3. aktual. u. erw. Auflage. Pearson, München.

Büttner H., 2001. Wassermanagement und Ressourcenkonflikte. Verlag für Entwicklungspolitik, Saarbrücken.

Chedid L.G., 2005. Energy, Society, and Education, with Emphasis on Educational Technology Policy for K-12. Journal of Science Education and Technology, 14(1), 75–85, DOI: 10.1007/s10956–005–2735–0.

Chomsky N., 1968. Language and mind. Harcourt, Brace & World, New York.

Collins L.M., Schafer J.L., Kam C.M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6, 330–351.

de Haan G., Kamp G., Lerch A., Martignon L., Müller-Christ G., Nutzinger H.G., 2008. Nachhaltigkeit und Gerechtigkeit. Springer-Verlag, Berlin Heidelberg.

Deffner G., 1984. Lautes Denken – Untersuchung zur Qualität eines Datenerhebungsverfahrens. Lang, Frankfurt (Main).

Dempsey J., 2012. Merkel Pays a Price for Her Energy Policy Shift. The New York Times, Ausgabe vom 28.05.2012.

dena, 2012. Integration der erneuerbaren Energien in den deutschen/europäischen Strommarkt. Deutsche Energie-Agentur GmbH, online verfügbar unter: http://www.dena.de/fileadmin/user_upload/Publikationen/Energiesysteme/Dokumente/dena-VNS_Abschlussbericht.pdf, abgerufen am 01.04.2013.

DGFG, 2012. Bildungsstandards im Fach Geographie für den Mittleren Schulabschluss, Ausgabe Beschlüsse der Kultusministerkonferenz. Selbstverlag Deutsche Gesellschaft für Geographie (DGfG), online verfügbar unter http://www.geographie.de/docs/geographie_bildungsstandards.pdf , 7. Auflage. Im Internet abrufbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Physik-Mittleren-SA.pdf, abgerufen am 04.06.2009.

Duit R., 1986. Der Energiebegriff im Physikunterricht. Dissertation, Institut für Didaktik der Naturwissenschaften (IPN) an der Christian-Albrechts-Universität Kiel.

DWV, 2011. DWV Wasserstoff-Sicherheits-Kompendium. online verfügbar unter http://www.dwv-info.de/publikationen/2011/sicher.pdf, abgerufen am 01.04.2013, Deutscher Wasserstoff- und Brennstoffzellenverband.

EC, 1995. Richtlinie 95/12/EG der Kommission vom 23. Mai 1995 zur Durchführung der Richtlinie 92/75/EWG des Rates betreffend die Energieetikettierung für elektrische Haushaltswaschmaschinen, online verfügbar unter: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1995L0012:19970117:DE:PDF, abgerufen am 08.11.2012, European Commission. Amtsblatt der Europäischen Union (ABl. L 136 vom 21.6.1995), Richtlinie 95/12/EG der EU-Kommission.

EC, 2007a. The EU climate and energy package. online verfügbar unter http://ec.europa.eu/clima/policies/package/index_en.htm, abgerufen am 14.02.2013, European Commission.

EC, 2007b. Renewable Energy Road Map – Renewable energies in the 21st century: building a more sustainable future, Ausgabe COM(2006) 848 final. Commission of the European Communities. Online verfügbar unter http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2006:0848:FIN:EN:PDF, abgerufen am 20.07.2012.

EC, 2009. Verordnung (EG) Nr. 244/2009 der Kommission vom 18. März 2009 zur Durchführung der Richtlinie 2005/32/EG des Europäischen Parlaments und

des Rates im Hinblick auf die Festlegung von Anforderungen an die umweltgerechte Gestaltung von Haushaltslampen mit ungebündeltem Licht. Amtsblatt der Europäischen Union (L 76/3 vom 24.3.2009): online verfügbar unter http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:076:0003:0016:DE:PDF, abgerufen am 20.07.2012, European Commission.

EC, 2011. Energy Roadmap 2050. COM(2011) 885/2, online verfügbar unter: http://ec.europa.eu/energy/energy2020/roadmap/doc/com_2011_8852_en.pdf, abgerufen am 14.02.2013, European Commission.

Eggert S., 2008. Bewertungskompetenz für den Biologieunterricht – Vom Modell zur empirischen Überprüfung, online verfügbar: http://webdoc.sub.gwdg.de/diss/2008/eggert/eggert.pdf. Dissertation, Georg-August-Universität Göttingen.

Eggert S., Bögeholz S., 2006. Göttinger Modell der Bewertungskompetenz – Teilkompetenz „Bewerten, Entscheiden und Reflektieren“ für Gestaltungsaufgaben Nachhaltiger Entwicklung. Zeitschrift für Didaktik der Naturwissenschaften, 12, 177–197.

Eggert S., Bögeholz S., 2010. Students' use of decision making strategies with regard to socioscientific issues – an application of the Rasch partial credit model. Science Education, 94, 230–258.

Eggert S., Hößle C., 2006. Bewertungskompetenz im Biologieunterricht. Ein Überblick. Praxis der Naturwissenschaften – Biologie in der Schule, 55(1), 1–10.

Eilks I., Feierabend T., Hößle C., Höttecke D., Menthe J., Mrochen M., Oelgeklaus H., 2011. Der Klimawandel vor Gericht. Aulis Verlag.

Einhorn H.J., Hogarth R.M., 1975. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 13, 171–192.

Engström S., Gustafsson P., Niedderer H., 2011. Content for Teaching sustainable Energy systems in Physics at Upper Secondary School. International Journal of Science and Mathematics Education, 9, 1281–1304.

Ericsson K.A., Simon H.A., 1984. Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge, MA.

Fishburn P.C., 1974. Lexicographic orders, utilities and decision rules: A survey. Management Science, 20, 1442–1471.

Fleiss J.L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378–382.

Fleiss J.L., 1981. Statistical methods for rates and proportions. 2. Aufl., Wiley, New York.

Gambro J.S., Switzky H.N., 1999. Variables Associated With American High School Students' Knowledge of Environmental Issues Related to Energy and Pollution. The Journal of Environmental Education, 30(2), 15–22.

Gausmann E., Eggert S., Hasselhorn M., Watermann R., Bögeholz S., 2010. Wie verarbeiten Schüler/innen Sachinformationen in Problem- und Entscheidungssituationen Nachhaltiger Entwicklung? In: E. Klieme; D. Leutner; M. Kenk (Hrsg.), Kompetenzmodellierung: Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes, Seiten 204–215. Zeitschrift für Pädagogik, 56. Beiheft.

Gillhaus A., 2007. Natural Gas Storage in Salt Caverns – Present Status, Developments and Future Trends in Europe. Bedienungsanleitung, Solution Mining Research Institute – Technical Conference Paper – Spring 2007 Conference.

Gläser-Zikuda M., Seidel T., Rohlfs C., Gröschner A., Ziegelbauer S. (Hrsg.), 2012. Mixed Methods in der empirischen Bildungsforschung. Waxmann.

Goldstein D.G., Gigerenzer G., 2002. Models of ecological rationality: The recognition heuristic. Psychological Review, 109, 75–90.

Graham J., 2012. Missing data. Springer-Verlag, DOI: 10.1007/978-1-4614-4018-5.

Graham J.W., Cumsille P.E., Elek-Fisk E., 2003. Methods for handling missing data. In: J.A. Schinka; W.F. Velicer (Hrsg.), Handbook of psychology: Research methods in psychology, Ausgabe 2, Seiten 87–114. Wiley, New York.

Gresch H., Bögeholz S., 2013. Identifying Non-Sustainable Courses of Action: A Prerequisite for Decision-Making in Education for Sustainable Development. Research in Science Education, 43(2), 733–754.

Gresch H., Hasselhorn M., Bögeholz S., 2011. Training in Decision-making Strategies: An Approach to enhance students' competence to deal with socio-scientific issues. International Journal of Science Education, 35 (15), 2587–2607, DOI:10.1080/09500693.2011.617789.

Greve W., 1997. Wissenschaftliche Beobachtung. Psychologie Verlags Union, Weinheim.

Grober U., 1999. Der Erfinder der Nachhaltigkeit. Die Zeit, Ausgabe 48.

GWEC, 2012. Global Wind Statistics 2011. Global Wind Energy Council. http://www.gwec.net/fileadmin/images/News/Press/GWEC_-_Global_Wind_ Statistics_2011.pdf, abgerufen am 19.07.2012.

Haidt J., 2001. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. Psychological Review, 108 (4), 814–834.

Halverson K.L., Siegel M.A., Freyermuth S.K., 2009. Lenses for Framing Decisions: Undergraduates' decision making about stem cell research. International Journal of Science Education, 31(9), 1249–1268, DOI: 10.1080/09500690802178123.

Hartig G.L., 1791. Anweisung zur Holzzucht für Förster. Marburg.

Hartmann S., 2013. Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz. Dissertation, Fakultät für Bildungswissenschaften, Universität Duisburg-Essen.

Heitmann P., 2012. Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. In: H. Niedderer; H. Fischler; E. Sumfleth (Hrsg.), Studien zum Physik- und Chemielernen, Ausgabe 121 von Studien zum Physik- und Chemielernen. Logos Verlag, Berlin.

Hohensinn C., Kubinger K.D., 2011. On the impact of missing values on item fit and the model validness of the Rasch model. Psychological Test and Assessment Modeling, 53 (3), 380–393.

Hostenbach J., 2011. Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht. In: H. Niedderer; H. Fischler; E. Sumfleth (Hrsg.), Studien zum Physik- und Chemielernen, Ausgabe 121 von Studien zum Physik- und Chemielernen. Logos Verlag Berlin.

Hötker H., Thomsen K.M., Jeromin H., 2006. Impacts on biodiversity of exploitation of renewable energy sources: the example of birds and bats - facts, gaps in knowledge, demands for further research, and ornithological guidelines for the development of renewable energy exploitation. http://bergenhusen.nabu.de/bericht/englische%20windkraftstudie.pdf, abgerufen am 19.07.2012, Michael-Otto-Institut im NABU, Bergenhusen.

Höttecke D., Mrochen M., 2010. Bewerten Lernen im Treibhaus. Praxis der Naturwissenschaften – Physik in der Schule, 59(2), 26–35.

IEA, 2012. World Ernergy Outlook 2012. online verfügbar unter: http://www.worldenergyoutlook.org/publications/weo-2012/, abgerufen am 12.02.2013, International Energy Agency.

Jimenez-Aleixandre M.P., 2002. Knowledge Producers or Knowledge Consumers? Argumentation and Decision making about environmental management. International Journal of Science Education, 24(11), 1171–1190.

Jungermann H., Pfister H.R., Fischer K., 2005. Die Psychologie der Entscheidung, 2. Auflage. Elsevier, München.

Kauertz A., 2007. Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben. In: H. Niedderer; H. Fischler; E. Sumfleth (Hrsg.), Studien zum Physik- und Chemielernen, Ausgabe 107 von Studien zum Physik- und Chemielernen. Logos Verlag, Berlin.

Kauertz A., Fischer H.E., Mayer J., Sumfleth E., Walpuski M., 2012. Standardbezogene Kompetenzmodelliereung in den Naturwissenschaften der Sekundarstufe I. Zeitschrift für Didaktik der Naturwissenschaften, 16, 135–153.

Klieme E., Avenarius H., Blum W., Döbrich P., Gruber H., Prenzel M., 2003. Zur Entwicklung nationaler Bildungsstandards – Eine Expertise. Bundesministerium für Bildung und Forschung, online verfügbar unter http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf, abgerufen am 03.09.2012.

Klieme E., Hartig J., 2008. Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. Zeitschrift für Erziehungswissenschaft, 10. Jg., 8. Sonderheft, 11–29.

KMK, 2004. Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss – Beschluss vom 04.12.2003, Ausgabe Beschlüsse der Kultusministerkonferenz. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Luchterhand. Online verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf, abgerufen am 04.06.2009.

KMK, 2005a. Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss – Beschluss vom 16.12.2004, Ausgabe Beschlüsse der Kultusministerkonferenz. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Luchterhand. Online verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf, abgerufen am 04.06.2009.

KMK, 2005b. Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss – Beschluss vom 16.12.2004, Ausgabe Beschlüsse der Kultusministerkonferenz. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Luchterhand. Online verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Chemie.pdf, abgerufen am 04.06.2009.

KMK, 2005c. Bildungsstandards im Fach Physik für den Mittleren Schulabschluss – Beschluss vom 16.12.2004, Ausgabe Beschlüsse der Kultusministerkonferenz. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Luchterhand. Online verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Physik-Mittleren-SA.pdf, abgerufen am 04.06.2009.

KMK, 2005d. Erläuterungen zur Konzeption und Entwicklung – Erläuterungen zur Konzeption und Entwicklung. online verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Konzeption-Entwicklung.pdf, abgerufen am 07.11.2012, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Luchterhand.

Knittel C., Mikelskis-Seifert S., 2011a. Erhebung von Bewertungskompetenz mittels Fragebogen?! online verfügbar unter http://phydid.physik.fu-berlin.de/index.php/phydid-b/article/view/246, abgerufen am 17.10.2012, Phydid-B, Beitrag DD 09.01.

Knittel C., Mikelskis-Seifert S., 2011b. Förderung von Bewertungskompetenz von Schülerinnen und Schülern durch Unterricht zur Photovoltaik. In: D. Höttecke (Hrsg.), Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie. Jahrestagung der GDCP in Potsdam 2010, LIT-Verlag, Berlin.

Knittel C., Mikelskis-Seifert S., 2012. Wie lernen Schülerinnen und Schüler bewerten? In: S. Bernholt (Hrsg.), Konzepte fachdidaktischer Strukturierung für den Unterricht. Jahrestagung der GDCP in Oldenburg 2011, LIT-Verlag, Berlin.

Knittel C., Mikelskis-Seifert S., 2013. Wie verändert Unterricht das Bewerten? - Messinstrumente / Modelle / Erkenntnisse. In: S. Bernholt (Hrsg.), Inquiry-based Learning - Forschendes Lernen. Jahrestagung der GDCP in Hannover 2012, LIT-Verlag, Berlin.

Kolstø S.D., 2006. Patterns in students' argumentation confronted with a risk-focused socio-scientific issue. International Journal of Science Education, 28(14), 1689–1716.

Lachmayer S., 2008. Entwicklung und Überprüfung eines Strukturmodells der Diagrammkompetenz für den Biologieunterricht. Dissertation, Mathematisch-Naturwissenschaftliche Fakultät, Chritian-Albrechts-Universität Kiel.

Leonhard W., Buenger U., Crotogino F., Gatzen C., Glaunsinger W., Huebner S., Kleimaier M., Koenemund M., Landinger H., Lebioda T., Sauer D.U., Weber H., Wenzel A., Wolf E., Woyke W., Zunft S., 2008. Energiespeicher in Stromversorgungssystemen mit hohem Anteil erneuerbarer Energieträger. Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE), Frankfurt am Main.

Liarakou G., Gavrilakis C., Flouri E., 2009. Secondary School Teachers' Knowledge and Attitudes Towards Renewable Energy Sources. Journal of Science Education and Technology, 18, 120–129.

Linacre J.M., 1998. Detecting Multidimensionality: Which Residual Data-type Works Best? Journal of Outcome Measurement, 2(3), 266–283.

Liu X., McKeogh A., 2005. Developement Growth in Students' Concept of Energy: Analysis of Selected Items from the TIMSS Database. Journal of Research in Science Teaching, 42(5), 493–517.

Lüdtke O., Robitzsch A., Trautwein U., Köller O., 2007. Umgang mit fehlenden Werten in der psychologischen Forschung – Probleme und Lösungen. Psychologische Rundschau, 58(2), 103–117.

Mannel S., 2010. Assessing scientific inquiry. Developement and evaluation of a test for the low-performing stage. In: H. Niedderer; H. Fischler; E. Sumfleth (Hrsg.), Studien zum Physik- und Chemielernen, Ausgabe 111 von Studien zum Physik- und Chemielernen. Logos Verlag, Berlin.

Martin H., Pohl R., 1996. Talsperren – nötig, sicher, ökologisch? Betrachtungen zur Überflutungssicherheit. Allianz-Report für Risiko und Sicherheit, 69(2), 64–68.

Masters G.N., 1982. A Rasch Model for Partial Credit Scoring. Psychometrika, 47(2), 149–172.

Mayer J., Harms U., Hammann M., Bayrhuber H., Kattmann U., 2004. Kerncurriculum Biologie der gymnasialen Oberstufe. MNU, 57/3, 166–173.

Mayring P., 2008. Qualitative Inhaltsanalyse. Beltz-Verlag, Weinheim.

Menthe J., 2012. Wider besseren Wissens?! Zeitschrift für interpretative Schul- und Unterrichtsforschung, 1, 161–183.

Menthe J., Düker P., 2013. Wie Subjekte urteilen und entscheiden – Habitus, Intuitionen, implizites Wissen und bewährte Strategien. In: S. Bernholt (Hrsg.), Inquiry-based Learning – Forschendes Lernen. Tagungsband der GDCP-Jahrestagung 2012, LIT-Verlag, Berlin.

Meyer R., 2010. Bewertungskompetenz im Physikunterricht – Eine qualitative Studie mit der Methode des Lauten Denkens. Staatsexamensarbeit, Didaktik der Physik, Georg-August-Universität Göttingen.

Mikelskis H., 1979. Zum Verhältnis von Wissenschaft und Lebenswelt im Physikunterricht – Darestellt am Thema Kernkraftwerke. Dissertation, Universität Bremen.

Mittelsten Scheid N., Hößle C., 2008. Bewerten im Biologieunterricht: Niveaus von Bewertungskompetenz. Erkenntnisweg Biologiedidaktik, 6, 87–104.

Moosbrugger H., Kelava A., 2008. Testtheorie und Fragebogenkonstruktion. Springer-Verlag, Heidelberg.

NdsMK, 2007. Kerncurriculum für das Gymnasium Schuljahrgänge 5-10 – Naturwissenschaften. online verfügbar unter: http://db2.nibis.de/1db/cuvo/datei/kc_gym_nws_07_nib.pdf, abgerufen am 19.07.2012, Niedersächisches Kultusministerium, unidruck, Hannover.

NdsMK, 2009. Kerncurriculum für das Gymnasium – gymnasiale Oberstufe – Physik. online verfügbar unter: http://db2.nibis.de/1db/cuvo/datei/kc_physik_go_i_2009.pdf, abgerufen am 19.07.2012, Niedersächisches Kultusministerium, unidruck, Hannover.

NdsMK, 2010. Kerncurriculum für das Gymnasium – gymnasiale Oberstufe – Erdkunde. online verfügbar unter: http://db2.nibis.de/1db/cuvo/datei/kc_erdkunde_go_i_03-11.pdf, abgerufen am 19.07.2012, Niedersächisches Kultusministerium.

Neumann K., Viering T., Boone W.J., Fischer H.E., 2013. Towards a learning progression of energy. Journal of Research in Science Teaching, 50(2), 162–188.

Newell A., Simon H.A., 1972. Human problem solving. Prentice-Hall, Englewood Cliffs, NJ.

Newman D.A., 2003. Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. Organizational Research Methods, 6, 328–362.

NRWSW, 2008. Kernlehrplan für das Gymnasium – Sekunderstufe I in Nordrhein-Westfalen. online verfügbar unter: http://www.schulentwicklung.nrw.de/lehrplaene/upload/lehrplaene_download/gymnasium_g8/gym8_physik.pdf, abgerufen am

19.07.2012, Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, Ritterbach Verlag, Frechen.

OECD, 2002. PISA 2000 Technical Report. Organisation for Economic Co-operation and Development, OECD Publishing.

OECD, 2005a. Definition und Auswahl von Schlüsselkompetenzen – Zusammenfassung. Organisation for Economic Co-operation and Development.

OECD, 2005b. PISA 2003 Technical Report. Organisation for Economic Co-operation and Development, OECD Publishing.

OECD, 2009. PISA 2006 Technical Report. Organisation for Economic Co-operation and Development, OECD Publishing.

Pahl E.M., Peters S., Komorek M., 2010. energie.bildung – Physik im Kontext von „Energiebildung“. PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung, online verfügbar unter: http://www.phydid.de/index.php/phydid-b/article/view/166/174, abgerufen am 15.08.2013.

Payne J.W., Bettman J.R., Johnson E.J., 1988. Adaptive strategy selection in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 534–552.

Payne J.W., Bettman J.R., Johnson E.J., 1993. The adaptive decision maker. Cambridge University Press, Cambridge.

Pedersen E., Hallberg L.M., Waye K.P., 2007. Living in the Vincinity of Wind Turbines – A Grounded Theory Study. Qualitative Research in Psychology, 4, 49–63.

Pospeschill M., 2010. Testtheorie, Testkonstruktion, Testevaluation. Reinhardt Verlag, München.

Prenzel M., Baumert J., Blum W., Lehmann R., Leutner D., Neubrand M., Pekrun R., Rost J., Schiefele U., 2006. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres. Waxmann.

Prüfer P., Rexroth M., 2005. Kognitive Interviews. ZUMA How-to-Reihe, Nr. 15, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, online verfügbar: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf, abgerufen am 22.10.2012.

Quaschning V., 2011. Regenerative Energiesysteme. Hanser Verlag, München.

Ratcliffe M., 1997. Pupil decision-making about socio-scientific issues within the science curriculum. International Journal of Science Education, 19(2), 167–182.

Ratcliffe M., Grace M., 2003. Science Education for Citizenship. Open University Press, Maidenhead.

Robottom I., Simonneaux L., 2012. Editorial: Socio-Scientific Issues and Education for Sustainability in Contemporary Education. Research in Science Education, 42, 1–4, DOI: 10.1007/s11165–011–9253–2.

Ropohl M., 2010. Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. Entwicklung und Analyse von Testaufgaben. In: H. Niedderer; H. Fischler; E. Sumfleth (Hrsg.), Studien zum Physik- und Chemielernen, Ausgabe 107 von Studien zum Physik- und Chemielernen. Logos Verlag, Berlin.

Rose S.L., Barton A.C., 2012. Should Great Lakes City Build a New Power Plant? How Youth navigate Socioscientific Issues. Journal of Research in Science Teaching, 49(5), 541–567.

Rost D.H., 2005. Interpretation und Bewertung pädagogischer-psychologischer Studien. Beltz Verlag, Weinheim und Basel.

Rost J., 1999. Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau, 50 (3), 140–156.

Rost J., 2002. Umweltbildung – Bildung für eine nachhaltige Entwicklung: Was macht den Unterschied? online verfügbar unter: http://satgeo-muenchen.de/geomuc/infos/umweltbildung.pdf, abgerufen am 24.10.2012.

Rost J., 2004. Lehrbuch Testtheorie – Testkonstruktion, 2. vollständig überarbeitete und erweiterte Auflage. Verlag Hans Huber, Bern.

Rost J., von Davier M., 1994. A Conditional Item-Fit Index for Rasch Models. Applied Psychological Measurement, 18, 171–182.

Rubin D.B., 1987. Multiple imputation for nonresponse in surveys. Wiley, New York.

Sadler T., 2004. Informal reasoning regarding socioscientific issues: A critical review of research. Journal of Research in Science Teaching, 41, 513–536.

Sadler T.D., 2011. Socio-Scientific Issues in the Classroom. Springer-Verlag, New York.

Sakschewski M., Bögeholz S., Eggert S., Meyer R., Schneider S., 2013a. Bewertungskompetenz im Physikunterricht: Erste Ergebnisse einer Studie mit Lautem Denken. In: S. Bernholt (Hrsg.), Inquiry-based Learning - Forschendes Lernen. Jahrestagung der GDCP

in Hannover 2012 [Beitrag zur Jahrestagung der GDCP in Potsdam 2010], 146-148, online verfügbar unter: http://www.gdcp.de/index.php/tagungsbaende/tagungsband-uebersicht/145-tagungsbaende/2013/4220-band33, 146-148, abgerufen am 02.04.2013.

Sakschewski M., Bögeholz S., Eggert S., Schneider S., 2013b. Messinstrument zum Basiskonzept Energie für Bewertungskompetenz: Ergebnisse zum Bewerten, Entscheiden und Reflektieren. In: S. Bernholt (Hrsg.), Inquiry-based Learning - Forschendes Lernen. Jahrestagung der GDCP in Hannover 2012, online verfügbar unter: http://www.gdcp.de/index.php/tagungsbaende/tagungsband-uebersicht/145-tagungsbaende/2013/4220-band33, 149-151, abgerufen am 02.04.2013.

Sauer D.U., 2010. Technologien, Einsatzszenarien und Kosten von Speichern für elektrische Energie. Vortrag, online verfügbar unter: http://www.ea-aw.de/fileadmin/downloads/Tagungen/FT10_Vortrag_Sauer.pdf, abgerufen am 07.11.2012.

Schafer J.L., 1997. Analysis of incomplete multivariate data. Chapman & Hall: London.

Schecker H., Höttecke D., 2007. Aufgaben zum Kompetenzbereich Bewerten. Naturwissenschaften im Unterricht Physik, 18(97), 29–36.

Schneider W., Schlagmüller M., Ennemoser M., 2007. Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12. Hogrefe, Göttingen.

Schott F., Azizi Ghanbari S., 2012. Bildungsstandards, Kompetenzdiagnostik und kompetenzorientierter Unterricht zur Qualitätssicherung des Bildungswesens. Waxmann.

Seethaler S., Linn M., 2004. Genetically modified food in perspective: An inquiry-based curriculum to help middle school students make sense of tradeoffs. International Journal of Science Education, 26(14), 1765–1785.

Shin S.H., 2009. How to treat omitted responses in Rasch Model-Based Equating. Practical Assessment, Research & Evaluation, 14(1).

Sill H.D., 2004. Bemerkungen zu den aktuellen Bildungsstandards. GDM-Mitteilungen, Nr. 78, 72–75.

Simon H.A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics, 69, 99–118.

Simonneaux J., Simonneaux L., 2012. Educational Condigurations for Teaching Environmental Socioscientific Issues Within the Perspective of Sustainability. Research in Science Education, 42, 75–94, DOI: 10.1007/s11165–011–9262–1.

Sinharay S., Stern H., Russell D., 2001. The use of multiple imputation for the analysis of missing data. Psychological Methods, 6, 317–329.

SRU, 1994. Umweltgutachten des Sachverständigenrats für Umweltfragen – Für eine dauerhaft-umweltgerechte Entwicklung. Bundestags-Drucksache 12/6995, Sachverständigenrat für Umweltfragen.

Strobl C., 2010. Das Rasch-Modell. Rainer Hampp Verlag, München und Mering.

Thorngate W., 1980. Efficient decision heuristics. Behavioral Science, 25, 219–225.

Tversky A., 1972. Elimination by aspects: A theory of choice. Psychological Review, 79, 281–299.

Tytler R., 2012. Socio-Scientific Issues, Sustainability and Science Education. Research in Science Education, 42, 155–163, DOI: 10.1007/s11165–011–9262–1.

UBA, 2007. Zukunftsmarkt Elektrische Energiespeicherung. online verfügbar unter: http://www.umweltdaten.de/publikationen/fpdf-l/3448.pdf, abgerufen am 18.02.1013, Umweltbundesamt.

UNCED, 1992. Agenda 21. online verfügbar unter: http://www.un.org/Depts/german/conf/agenda21/agenda_21.pdf, abgerufen am 13.02.2013, United Nations Conference on Environment and Development.

UNESCO, 2005. United Nations Decade of Education for Sustsainable Development (2005-2014): International Implementation Scheme. Section for Education for Sustainable Development (ED/PEQ/ESD), online verfügbar unter: http://unesdoc.unesco.org/images/0014/001486/148654e.pdf, abgerufen am 20.04.2013, United Nations Educational, Scientific and Cultural Organization.

Uskola A., Maguregia G., Jiménez-Aleixandre M.P., 2010. The Use of Criteria in Argumentation and the Construction of Environmental Concepts: A university case study. International Journal of Science Education, 32(17), 2311–2333.

Vaughn N., 2009. Wind energy – renewable energy and the environment. CRC Press, Taylor & Francis Group.

VDE, 2009. Energiespeicher in Stromversorgungssystemen mit hohem Anteil erneuerbarer Energieträger – Bedeutung, Stand der Technik, Handlungsbedarf. Energietechnische Gesellschaft im VDE (ETG), Verband Deutscher Elektroingenieure.

Viering T., 2012. Entwicklung physikalischer Kompetenz in der Sekundarstufe I. In: H. Niedderer; H. Fischler; E. Sumfleth (Hrsg.), Studien zum Physik- und Chemielernen, Ausgabe 121 von Studien zum Physik- und Chemielernen. Logos Verlag, Berlin.

Walpuski M., Kampa N., Kauertz A., Wellnitz N., 2008. Evaluation der Bildungsstandards in den Naturwissenschaften. MNU, 61 (6), 323–326.

Walpuski M., Kauertz A., Fischer H.E., Kampa N., Mayer J., Sumfleth E., Wellnitz N., 2010. ESNaS – Evaluation der Standards für die Naturwissenschaften in der Sekundarstufe I. In: A. Gehrmann (Hrsg.), Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht. Klinkhardt, Bad Heilbrunn.

Walpuski M., Sumfleth E., 2010. Überprüfung der nationalen Bildungsstandards im Fach Chemie. Praxis der Naturwissenschaften – Chemie in der Schule, 33 (52), 24–28.

WCED, 1987. Our common future. Report of the World Commission on Environment and Development. Oxford, online verfügbar unter: http://www.un-documents.net/our-common-future.pdf, abgerufen am 23.08.2013, World Commission on Environment and Development.

Weinert F.E., 2001. Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: Leistungsmessungen in Schulen, S. 17-31. Beltz-Verlag, Weinheim und Basel.

WEMAG, 2013. Größter Batteriespeicher Europas entsteht in Schwerin. online verfügbar unter: https://www.wemagblog.com/2013/04/29/groster-batteriespeicher-europas-entsteht-in-schwerin/ zuletzt abgerufen am 04.09.2013.

Wirtz M., Caspar F., 2002. Beurteilerübereinstimmung und Beurteilerreliabilität. Hogrefe, Göttingen.

Wolsink M., 2007. Wind power implementation: The nature of public attitudes: Equity and fairness instead of 'backyard motives'. Renewable and Sustainable Energy Reviews, 11, 118–1207.

Wu M., Adams R., 2007. Applying the Rasch model to psycho-social measurement: A practical approach. Educational Measurement Solutions, Melbourne.

Wu M.L., Adams R.J., Wilson M.R., Haldane S.A., 2007. ACER ConQuest version 2.0: generalised item response modelling software. ACER Press.

WWEA, 2011. World Wind Energy Report 2010. online verfügbar unter: http://www.wwindea.org/home/images/stories/pdfs/worldwindenergyreport2010_s.pdf, abgerufen am 23.11.2012, World Wind Energy Association.

Zöfel P., 2003. Statistik für Psychologen – im Klartext. Pearson Studium, München.

ANHANG A

Anhang

A.1 Fit-Statistiken auf Basis der multiplen Imputationen

Auf den folgenden Seiten sind die Ergebnisse der fünf beispielhaft imputierten Datensätze abgedruckt. Die Item-Fit-Indizes des Original-Datensatzes sind in der Tabelle 6.7 enthalten.

Das Verfahren der multiplen Imputation kam zur Anwendung, um eine systematische Verzerrung der Ergebnisse auszuschließen, die durch listenweisen Ausschluss zustande kommen könnte (siehe Kapitel 5.2).

Tabelle A.1: Item-Fit-Indizes eines mittels multipler Imputation vervollständigten Datensatzes (Imputation Nr. 1) mit den Programmen WINMIRA und ConQuest beim Rasch-Partial-Credit-Modell.

Item-Fit-Indizes																		
	Itemnummer:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Kürzel[a]	Opt	Opt	Opt	Opt	Gew	Opt	Opt	Opt	Opt	Gew	MCi	MCn	MCk	PE	VSn	VSk	VSi
Parameter[b]	Vorgabe	Aufgabe 1					Aufgabe 2					Aufgabe 3						
Item-Fit-Indizes mit WINMIRA (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
Q-Index	$< 0{,}35$	0,22	0,25	0,19	0,17	0,29	0,18	0,21	0,23	0,22	0,30	0,23	0,25	0,32	0,28	0,24	0,30	0,25
Z_Q	$\in [-1{,}96; 1{,}96]$	-1,05	0,11	-2,00^{O}	-3,19^{O}	1,44	-3,01^{O}	-1,63	-0,43	-0,90	1,72	-0,59	0,05	2,93^{U}	0,11	2,22^{U}	2,22^{U}	0,49
Z_Q	sign.?	–	–	*	**	–	**	–	–	–	*	–	–	**	–	*	*	–
Thresholds	$\in [-3; 3]$	0,16	-0,04	-0,29	-0,12	0,86	1,16	0,48	-0,28	-0,57	1,26	-1,65	-1,23	0,41	-2,61	-0,13	0,64	1,94
Item-Fit-Indizes mit ConQuest (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
WMNSQ	$\in [0{,}75; 1{,}33]$	0,97	1,00	0,94	0,90	1,02	0,92	0,98	0,98	0,98	1,03	0,97	0,99	1,07	1,01	1,07	1,07	1,00
dazu T-Wert	$\in [-2; 2]$	-1,2	0,1	-2,9^{O}	-4,3^{O}	0,5	-1,6	-0,7	-0,9	-1,0	0,6	-0,7	-0,3	2,1^{U}	0,2	1,8	1,6	0,0
Trennschärfe	$>0{,}25$	0,44	0,40	0,50	0,54	0,30	0,45	0,45	0,43	0,45	0,26	0,39	0,38	0,28	0,26	0,47	0,33	0,34
Item-Deltas[c]	$\in [-3; 3]$	0,61	0,41	0,15	0,33	1,30	1,61	0,93	0,17	-0,12	1,70	-1,20	-0,79	0,86	-2,17	0,31	1,09	2,40

Item-Fit-Indizes für polytome Items																
		WINMIRA			ConQuest											
		Thresholds			Item-Thresholds			Item-Deltas			Pt Bis[d]			PV1Avg:1		
		für Score...			für Score...			für Score...			für Score...			für Score...		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Item 15	VSn	–	-0,67	0,41	–	-0,46	1,09	–	-0,22	0,85	-0,40	0,05	0,40	-0,34	0,04	0,47
Item 16	VSk	–	-1,21	2,50	–	-0,79	2,97	–	-0,77	2,95	-0,30	0,21	0,19^{N}	-0,28	0,10	0,55
Item 17	VSi	–	1,32	2,55	–	1,55	3,23^{S}	–	1,76	3,03^{S}	-0,33	0,30	0,15^{N}	-0,10	0,41	0,64

[a] Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtfertigendem Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtfertigendes Vorgehen, non-kompensatorischer und kompensatorischer Strategie

[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.

[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.

[d] Alle Pt Bis-Korrelationen sind signifikant mit $p<0{,}001$, bis auf die für Score 1 bei Item 15 (nicht signifikant).

[O, U] Leichter Overfit bzw. Underfit eines Items

[S] Item ist geringfügig zu schwierig, da Item-Threshold bzw. Item-Delta $>3{,}00$.

[N] Score 2 bei diesem Item ist nicht geordnet, aber positiv mit Gesamtskala korreliert.

Tabelle A.2: Item-Fit-Indizes eines mittels multipler Imputation vervollständigten Datensatzes (Imputation Nr. 2) mit den Programmen WINMIRA und ConQuest beim Rasch-Partial-Credit-Modell.

Item-Fit-Indizes																		
	Itemnummer:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Kürzel[a]	Opt	Opt	Opt	Opt	Gew	Opt	Opt	Opt	Opt	Gew	MCi	MCn	MCk	PE	VSn	VSk	VSi
Parameter[b]	Vorgabe	Aufgabe 1					Aufgabe 2					Aufgabe 3						
Item-Fit-Indizes mit WINMIRA (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
Q-Index	$< 0{,}35$	0,22	0,25	0,19	0,17	0,29	0,17	0,21	0,23	0,22	0,30	0,23	0,25	0,31	0,28	0,23	0,31	0,25
Z_Q	$\in [-1{,}96; 1{,}96]$	-1,02	0,24	-2,07^{O}	-3,20^{O}	1,39	-3,23^{O}	-1,64	-0,55	-0,87	1,79	-0,56	0,12	2,85^{U}	0,01	2,07^{U}	2,53^{U}	0,62
Z_Q	sign.?	–	–	*	**	–	**	–	–	–	*	–	–	**	–	*	**	–
Thresholds	$\in [-3; 3]$	0,16	-0,04	-0,29	-0,12	0,86	1,15	0,47	-0,29	-0,57	1,26	-1,66	-1,23	0,41	-2,62	-0,13	0,69	1,96
Item-Fit-Indizes mit ConQuest (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
WMNSQ	$\in [0{,}75; 1{,}33]$	0,97	1,00	0,93	0,90	1,02	0,92	0,97	0,98	0,97	1,03	0,98	0,99	1,06	1,03	1,05	1,09	0,99
dazu T-Wert	$\in [-2; 2]$	-1,3	0,1	-3,0^{O}	-4,2^{O}	0,4	-1,5	-0,9	-0,7	-1,3	0,6	-0,5	-0,3	1,9	0,5	1,3	1,9	-0,1
Trennschärfe	$>0{,}25$	0,44	0,40	0,50	0,54	0,30	0,45	0,45	0,44	0,45	0,26	0,39	0,38	0,28	0,27	0,47	0,32	0,34
Item-Deltas[c]	$\in [-3; 3]$	0,61	0,41	0,16	0,33	1,31	1,60	0,92	0,17	-0,12	1,71	-1,20	-0,78	0,86	-2,17	0,32	1,15	2,43

Item-Fit-Indizes für polytome Items																
		WINMIRA			ConQuest											
		Thresholds für Score...			Item-Thresholds für Score...			Item-Deltas für Score...			Pt Bis[d] für Score...			PV1Avg:1 für Score...		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Item 15	VSn	–	-0,69	0,43	–	-0,47	1,11	–	-0,23	0,87	-0,40	0,05	0,40	-0,33	0,04	0,46
Item 16	VSk	–	-1,18	2,56	–	-0,75	3,04^{S}	–	-0,72	3,02^{S}	-0,29	0,21	0,16^{N}	-0,26	0,10	0,49
Item 17	VSi	–	1,26	2,66	–	1,53	3,33^{S}	–	1,71	3,15^{S}	-0,33	0,30	0,13^{N}	-0,10	0,41	0,58

[a] Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtfertigendem Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtfertigendes Vorgehen, non-kompensatorischer und kompensatorischer Strategie

[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.

[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.

[d] Alle Pt Bis-Korrelationen sind signifikant mit $p<0{,}001$, bis auf die für Score 1 bei Item 15 (nicht signifikant).

[O, U] Leichter Overfit bzw. Underfit eines Items

[S] Item ist geringfügig zu schwierig, da Item-Threshold bzw. Item-Delta $>3{,}00$.

[N] Score 2 bei diesem Item ist nicht geordnet, aber positiv mit Gesamtskala korreliert.

Tabelle A.3: Item-Fit-Indizes eines mittels multipler Imputation vervollständigten Datensatzes (Imputation Nr. 3) mit den Programmen WINMIRA und ConQuest beim Rasch-Partial-Credit-Modell.

Item-Fit-Indizes																		
	Itemnummer:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Kürzel[a]	Opt	Opt	Opt	Opt	Gew	Opt	Opt	Opt	Opt	Gew	MCi	MCn	MCk	PE	VSn	VSk	VSi
Parameter[b]	Vorgabe	Aufgabe 1					Aufgabe 2					Aufgabe 3						
Item-Fit-Indizes mit WINMIRA (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
Q-Index	$< 0{,}35$	0,21	0,24	0,19	0,16	0,29	0,17	0,21	0,23	0,22	0,30	0,24	0,26	0,32	0,29	0,23	0,30	0,25
Z_Q	$\in [-1{,}96; 1{,}96]$	-1,07	0,06	-2,02^O	-3,25^O	1,38	-3,12^O	-1,68	-0,50	-0,98	1,75	-0,49	0,24	2,91^U	0,29	2,06^U	2,24^U	0,65
Z_Q	sign.?	–	–	*	**	–	**	*	–	–	*	–	–	**	–	*	*	–
Thresholds	$\in [-3; 3]$	0,15	-0,05	-0,29	-0,12	0,86	1,16	0,48	-0,29	-0,57	1,25	-1,65	-1,24	0,41	-2,65	-0,14	0,69	2,01
Item-Fit-Indizes mit ConQuest (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
WMNSQ	$\in [0{,}75; 1{,}33]$	0,97	1,01	0,93	0,91	1,02	0,92	0,97	0,98	0,98	1,03	0,99	1,00	1,05	1,03	1,06	1,07	0,99
dazu T-Wert	$\in [-2; 2]$	-0,9	0,2	-3,0^O	-4,0^O	0,4	-1,6	-0,9	-0,8	-0,9	0,5	-0,3	0,0	1,6	0,4	1,5	1,6	-0,1
Trennschärfe	$>0{,}25$	0,44	0,40	0,50	0,54	0,30	0,45	0,45	0,44	0,46	0,26	0,38	0,37	0,28	0,25	0,48	0,32	0,33
Item-Deltas[c]	$\in [-3; 3]$	0,60	0,40	0,16	0,33	1,31	1,61	0,93	0,17	-0,12	1,70	-1,20	-0,79	0,86	-2,21	0,31	1,15	2,48

Item-Fit-Indizes für polytome Items																
		WINMIRA			ConQuest											
		Thresholds			Item-Thresholds			Item-Deltas			Pt Bis[d]			PV1Avg:1		
		für Score...			für Score...			für Score...			für Score...			für Score...		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Item 15	VSn	–	-0,70	0,41	–	-0,48	1,09	–	-0,24	0,85	-0,40	0,04	0,41	-0,32	0,03	0,47
Item 16	VSk	–	-1,21	2,59	–	-0,77	3,08^S	–	-0,75	3,05^S	-0,30	0,22	0,17^N	-0,27	0,11	0,50
Item 17	VSi	–	1,28	2,73	–	1,55	3,39^S	–	1,73	3,22^S	-0,32	0,29	0,13^N	-0,09	0,40	0,61

[a] Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtfertigendem Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtfertigendes Vorgehen, non-kompensatorischer und kompensatorischer Strategie

[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.

[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.

[d] Alle Pt Bis-Korrelationen sind signifikant mit $p<0{,}001$, bis auf die für Score 1 bei Item 15 (nicht signifikant).

[O, U] Leichter Overfit bzw. Underfit eines Items

[S] Item ist geringfügig zu schwierig, da Item-Threshold bzw. Item-Delta $>3{,}00$.

[N] Score 2 bei diesem Item ist nicht geordnet, aber positiv mit Gesamtskala korreliert.

Tabelle A.4: Item-Fit-Indizes eines mittels multipler Imputation vervollständigten Datensatzes (Imputation Nr. 4) mit den Programmen WINMIRA und ConQuest beim Rasch-Partial-Credit-Modell.

Item-Fit-Indizes																		
	Itemnummer:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Kürzel[a]	Opt	Opt	Opt	Opt	Gew	Opt	Opt	Opt	Opt	Gew	MCi	MCn	MCk	PE	VSn	VSk	VSi
Parameter[b]	Vorgabe	Aufgabe 1					Aufgabe 2					Aufgabe 3						
Item-Fit-Indizes mit WINMIRA (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
Q-Index	$< 0{,}35$	0,22	0,24	0,19	0,17	0,28	0,17	0,21	0,23	0,22	0,30	0,23	0,25	0,32	0,27	0,23	0,31	0,25
Z_Q	$\in [-1{,}96; 1{,}96]$	-1,04	0,13	-1,98^{O}	-3,20^{O}	1,33	-3,09^{O}	-1,71	-0,40	-0,86	1,67	-0,54	0,25	3,06^{U}	-0,05	2,03^{U}	2,36^{U}	0,52
Z_Q	sign.?	–	–	*	**	–	**	*	–	–	*	–	–	**	–	*	**	–
Thresholds	$\in [-3; 3]$	0,15	-0,05	-0,29	-0,12	0,86	1,15	0,47	-0,28	-0,57	1,26	-1,67	-1,24	0,39	-2,51	-0,14	0,67	1,92
Item-Fit-Indizes mit ConQuest (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
WMNSQ	$\in [0{,}75; 1{,}33]$	0,97	1,00	0,95	0,91	1,01	0,93	0,98	0,99	0,97	1,02	0,97	0,99	1,07	1,01	1,04	1,08	1,00
dazu T-Wert	$\in [-2; 2]$	-1,1	-0,1	-2,5^{O}	-4,0^{O}	0,3	-1,4	-0,5	-0,3	-1,2	0,4	-0,7	-0,3	2,1^{U}	0,2	1,2	1,8	0,0
Trennschärfe	$>0{,}25$	0,44	0,40	0,50	0,54	0,30	0,45	0,45	0,43	0,45	0,26	0,39	0,37	0,27	0,29	0,48	0,32	0,33
Item-Deltas[c]	$\in [-3; 3]$	0,60	0,41	0,16	0,33	1,31	1,61	0,92	0,17	-0,12	1,71	-1,21	-0,78	0,84	-2,05	0,31	1,13	2,40

Item-Fit-Indizes für polytome Items																
		WINMIRA			ConQuest											
		Thresholds für Score...			Item-Thresholds für Score...			Item-Deltas für Score...			Pt Bis[d] für Score...			PV1Avg:1 für Score...		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Item 15	VSn	–	-0,69	0,41	–	-0,47	1,09	–	-0,23	0,86	-0,40	0,05	0,41	-0,32	0,06	0,49
Item 16	VSk	–	-1,22	2,57	–	-0,79	3,05^{S}	–	-0,77	3,03^{S}	-0,30	0,22	0,16^{N}	-0,23	0,11	0,53
Item 17	VSi	–	1,32	2,52	–	1,56	3,23^{S}	–	1,77	3,02^{S}	-0,33	0,30	0,14^{N}	-0,08	0,42	0,71

[a] Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtfertigendem Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtfertigendes Vorgehen, non-kompensatorischer und kompensatorischer Strategie

[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.

[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.

[d] Alle Pt Bis-Korrelationen sind signifikant mit $p<0{,}001$, bis auf die für Score 1 bei Item 15 (nicht signifikant).

[O, U] Leichter Overfit bzw. Underfit eines Items

[S] Item ist geringfügig zu schwierig, da Item-Threshold bzw. Item-Delta $>3{,}00$.

[N] Score 2 bei diesem Item ist nicht geordnet, aber positiv mit Gesamtskala korreliert.

Tabelle A.5: Item-Fit-Indizes eines mittels multipler Imputation vervollständigten Datensatzes (Imputation Nr. 5) mit den Programmen WINMIRA und ConQuest beim Rasch-Partial-Credit-Modell.

Item-Fit-Indizes																		
	Itemnummer:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Kürzel[a]	Opt	Opt	Opt	Opt	Gew	Opt	Opt	Opt	Opt	Gew	MCi	MCn	MCk	PE	VSn	VSk	VSi
Parameter[b]	Vorgabe	Aufgabe 1					Aufgabe 2					Aufgabe 3						
Item-Fit-Indizes mit WINMIRA (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
Q-Index	$< 0{,}35$	0,22	0,25	0,19	0,16	0,28	0,17	0,20	0,23	0,22	0,30	0,23	0,25	0,31	0,28	0,23	0,31	0,24
Z_Q	$\in [-1{,}96; 1{,}96]$	-1,00	0,24	-1,90	-3,21^{O}	1,25	-3,09^{O}	-1,80	-0,46	-0,91	1,68	-0,63	0,22	2,76^{U}	0,31	2,00^{U}	2,52^{U}	0,36
Z_Q	sign.?	–	–	*	**	–	**	*	–	–	*	–	–	**	–	*	**	–
Thresholds	$\in [-3; 3]$	0,14	-0,06	-0,31	-0,13	0,85	1,15	0,47	-0,29	-0,58	1,25	-1,66	-1,24	0,40	-2,53	-0,15	0,69	2,00
Item-Fit-Indizes mit ConQuest (N=850 mittels multipler Imputation vervollständigte Datensätze)																		
WMNSQ	$\in [0{,}75; 1{,}33]$	0,96	1,00	0,94	0,92	1,01	0,92	0,97	0,98	0,98	1,02	0,97	1,00	1,07	1,02	1,05	1,08	0,99
dazu T-Wert	$\in [-2; 2]$	-1,5	0,1	-2,5^{O}	-3,7^{O}	0,2	-1,5	-0,9	-0,8	-0,9	0,4	-0,6	0,1	2,2^{U}	0,4	1,3	1,9	-0,1
Trennschärfe	$>0{,}25$	0,44	0,40	0,50	0,54	0,31	0,45	0,45	0,43	0,45	0,26	0,39	0,38	0,28	0,27	0,48	0,31	0,33
Item-Deltas[c]	$\in [-3; 3]$	0,60	0,40	0,15	0,33	1,31	1,61	0,93	0,17	-0,12	1,71	-1,20	-0,77	0,86	-2,07	0,31	1,16	2,48

Item-Fit-Indizes für polytome Items																
		WINMIRA			ConQuest											
		Thresholds			Item-Thresholds			Item-Deltas			Pt Bis[d]			PV1Avg:1		
		für Score...			für Score...			für Score...			für Score...			für Score...		
		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Item 15	VSn	–	-0,70	0,40	–	-0,47	1,09	–	-0,23	0,86	-0,40	0,06	0,40	-0,31	0,06	0,44
Item 16	VSk	–	-1,19	2,57	–	-0,75	3,07^{S}	–	-0,72	3,05^{S}	-0,29	0,22	0,15^{N}	-0,24	0,11	0,49
Item 17	VSi	–	1,32	2,68	–	1,59	3,38^{S}	–	1,78	3,19^{S}	-0,33	0,30	0,13^{N}	-0,08	0,42	0,55

[a] Opt: Umgang mit Optionen; Gew: Gewichtung von Kriterien; MCi, MCn, MCk: Multiple-Choice-Fragen für Beschreibung von intuitiv-rechtfertigendem Vorgehen, non-kompensatorischer und kompensatorischer Strategie; VSi, VSn, VSk: Verbesserungsvorschlag für intuitiv-rechtfertigendes Vorgehen, non-kompensatorischer und kompensatorischer Strategie

[b] Die Parameter werden ausführlich in Kapitel 5.3.2 vorgestellt.

[c] Werden bei ConQuest auch als *Estimate* ausgegeben, hier erfolgt auch die Ausgabe der gemittelten Item-Deltas. Die Item-Deltas sind diejenigen Itemschwierigkeiten, die in der Person-Item-Map (Abbildung 6.2) aufgetragen sind.

[d] Alle Pt Bis-Korrelationen sind signifikant mit $p<0{,}001$, bis auf die für Score 1 bei Item 15 (nicht signifikant).

[O, U] Leichter Overfit bzw. Underfit eines Items

[S] Item ist geringfügig zu schwierig, da Item-Threshold bzw. Item-Delta $>3{,}00$.

[N] Score 2 bei diesem Item ist nicht geordnet, aber positiv mit Gesamtskala korreliert.

A.2 Verwendete Software

Für die Analyse des Datensatzes wurden die folgenden Softwarepakete verwendet:

- IBM SPSS Statistics Version 20
- ConQuest Version 2.0
- WINMIRA 2001 Version 1.45
- Microsoft Office Excel 2003
- MaxQDA 10

Ferner zur Verfassung der Dissertation unter anderem folgende Programme:

- TeXnicCenter 2.0 beta 1
- MiKTeX 2.9
- JabRef 2.8
- CorelDraw GraphicsSuite X3
- Notepad++ Version 6.4.3

A.3 Testinstrumentbezug und weitere Hinweise

Das Testinstrument ist aufgeteilt in ein Informationsheft und ein Antwortheft. Die inhaltlichen Überlegungen zur Ausgestaltung der Items und der Aufgabenkontexte sind in Kapitel 4 dargelegt. Das Testinstrument für die Schülerinnen und Schüler der Oberstufe wurde auf eine Form umgearbeitet, in der die Schülerinnen und Schüler gesiezt werden.

Das im Rahmen dieser Studie entwickelte und eingesetzte Testinstrument zur Messung und Modellierung von Bewertungskompetenz im Kontext nachhaltiger Entwicklung wird veröffentlicht beim *Forschungsdatenzentrum (FDZ) Bildung* und kann dort zusammen mit einem Testmanual angefragt werden.

Forschungsdaten Bildung:
Internetadresse: http://www.forschungsdaten-bildung.de

Veränderter Scoring Guide in dem beim FDZ Bildung veröffentlichten Testmanual

Hinweis: Bei dem beim FDZ Bildung veröffentlichten Testmanual kommt aus Gründen der Vergleichbarkeit mit früheren Studien zur Bewertungskompetenz im Biologieunterricht (Eggert 2008, Eggert & Bögeholz 2010) ein im Vergleich zu dieser Dissertation verändertes Scoring zum Einsatz. Die Variation bezieht sich auf das Item 16, bei dem eine Antwortkategorie anders zugeordnet wird.

Veröffentlichung im *International Journal of Science Education (IJSE)*

Im Rahmen des gesamten Projekts ist ein englischsprachiger Artikel entstanden, der mittlerweile beim IJSE veröffentlicht ist. Auch in dieser Veröffentlichung wurde der variierte Scoring Guide verwendet, weshalb die dort beschriebenen Item-Fit-Parameter und andere statistische Größen sich von denen in dieser Dissertation geringfügig unterscheiden.

Sakschewski, M., Eggert, S., Schneider, S., & Bögeholz, S., (2014). Students' Socioscientific Reasoning and Decision-making on Energy-related Issues – Development of a measurement instrument, International Journal of Science Education.
Online verfügbar unter: http://dx.doi.org/10.1080/09500693.2014.920550

Abbildungsverzeichnis

Tabellenverzeichnis

Danksagung

Während meiner Promotionszeit in der Didaktik der Physik der Fakultät für Physik bin ich vielen netten und hilfsbereiten Menschen begegnet, die mir ein sehr angenehmes und freundliches Arbeitsklima geboten haben. Ihnen und Euch allen dafür vorweg einen ganz herzlichen Dank!

Bei meiner Betreuerin Prof. Dr. Susanne Schneider bedanke ich mich ganz herzlich für die Gelegenheit, diese physikdidaktische Dissertation anfertigen zu können. Die langjährige liebevolle Zusammenarbeit wird mir in sehr guter Erinnerung bleiben.

Auch meiner anderen Betreuerin Prof. Dr. Susanne Bögeholz möchte ich ganz herzlich für die Kooperation in den zurückliegenden Jahren und die vielen wissenschaftlichen und methodischen Anregungen danken, die wesentlich zum Gelingen dieser Arbeit beigetragen haben.

Dies gilt ebenfalls für Dr. Sabina Eggert, die mit den Erfahrungen aus ihren eigenen Erhebungen stets wertvolle Informationen beisteuern und mir viele praktische Ratschläge geben konnte.

Meinem langjährigen Bürokollegen Arne Gerdes danke ich in ganz besonderer Weise für die vielen kritisch-konstruktiven Nachfragen und Hinweise in fachlicher Hinsicht und das freundschaftliche „Zusammenleben im Büro“ in den zurückliegenden Jahren.

Bei Rena Meyer bedanke ich mich einerseits ganz herzlich für ihre Mitarbeit bei Konzeption und Durchführung der ersten Vorstudie und andererseits zusammen mit den anderen Codiererinnen und Codierern Matthias Deters, Janine de Geus, Julia Horstmann, Hans-Ulrich Kaufeld und Michael Mathe für das Lesen und Verarbeiten der vielen Testmaterialien.

Schließlich bedanke ich mich bei den Schülerinnen und Schülern und den Lehrkräften der folgenden Gymnasien dafür, dass sie an Vorstudien und Hauptstudie so bereitwillig teilgenommen haben und damit im wahrsten Sinne des Wortes die Basis für diese

empirische Arbeit geliefert haben: Gymnasium Neue Oberschule Braunschweig, Otto-Hahn-Gymnasium Göttingen, Gymnasium Mellendorf und Gymnasium Walsrode, an dem ich damals selbst viele Jahre als Schüler verbracht habe.

Lebenslauf

Personalien	
Name:	Mark Torsten Sakschewski
Geburtstag:	08. August 1983
Staatsangehörigkeit:	deutsch
Adresse:	Heinefelder Weg 7
	33034 Brakel

Schulbildung	
08/1989-07/1993	Grundschule Süd, Walsrode
08/1993-07/1995	Orientierungsstufe Walsrode
08/1995-06/2002	Gymnasium Walsrode

Studium	
10/2002-08/2008	Physik (Diplom), Georg-August-Universität Göttingen
	Abschluss: 06.08.2008
04/2004-11/2008	Physik und Mathematik (Lehramt an Gymnasien, Erstes Staatsexamen),
	Georg-August-Universität Göttingen, Abschluss: 26.11.2008
04/2009-10/2013	Physik (Promotion), Georg-August-Universität Göttingen
	Disputation: 30.10.2013

Beruflicher Werdegang	
10/2004-09/2005	Studentische Hilfskraft (Physikpraktikum für Mediziner) IV. Physikalisches Institut, Georg-August-Universität Göttingen
05/2005-05/2011	Projektleiter Semesterticket Allgemeiner Studierendenausschuss (AStA) der Georg-August-Universität Göttingen
10/2006-05/2008	Studentische Hilfskraft (Geophysikpraktikum) Institut für Geophysik, Georg-August-Universität Göttingen
10/2008	Wissenschaftliche Hilfskraft Institut für Geophysik, Georg-August-Universität Göttingen
11/2008-04/2009	Wissenschaftlicher Mitarbeiter Institut für Geophysik, Georg-August-Universität Göttingen
02/2009-06/2009	Wissenschaftliche Hilfskraft IV. Physikalisches Institut, Didaktik der Physik, Georg-August-Universität Göttingen
07/2009-04/2013	Wissenschaftlicher Mitarbeiter IV. Physikalisches Institut, Didaktik der Physik, Georg-August-Universität Göttingen
05/2013-03/2014	Wissenschaftliche Hilfskraft IV. Physikalisches Institut, Didaktik der Physik, Georg-August-Universität Göttingen
05/2013-heute	Studienreferendar Zentrum für schulpraktische Lehrerausbildung Detmold

Tätigkeiten in der akademischen Selbstverwaltung	
04/2007-03/2009	Vorstand des Instituts für Geophysik, Georg-August-Universität Göttingen (Vertretung der Studierenden)
04/2010-03/2012	Vorstand des Zentrums für empirische Unterrichts- und Schulforschung (ZeUS), Georg-August-Universität Göttingen (Vertretung der wissenschaftlichen Mitarbeiter)

Göttingen, 25.09.2013