Sarah Gogolin & Dirk Krüger
Freie Universität Berlin

Konstruktion von Diagnoseaufgaben zum Zweck von Modellen

Development of diagnostic tasks for the purpose of models

Für die Entwicklung von Diagnoseinstrumenten wird in aktuellen Standards der Testentwicklung gefordert, die Schüler_innen stärker in den Prozess der Aufgabenentwicklung einzubeziehen. Dieser Beitrag schlägt eine auf Ratingaufgaben und Schülerinterviews basierende Prozedur zur Konstruktion von Forced Choice Aufgaben zum Zweck von Modellen vor. Die Konstruktionsschritte werden theoretisch begründet und an einem Beispiel illustriert. Die Prozedur ermöglichte es, sechs Diagnoseaufgaben zusammenzustellen, die eine theoriekonforme Interpretation des Modellverstehens von Schüler_innen ermöglichen.

Schlüsselwörter: Modelle, Biologie, Schülervorstellungen, Diagnose, Forced Choice Aufgaben, Validität

The latest standards for testing call for the integration of students into the process of test development. This article proposes a procedure for developing diagnostic tasks for the purpose of models which is both based on students' decisions on rating scales and on student interviews. The construction steps are being theoretically explained and justified as well as illustrated with an example. Using the procedure, six diagnostic tasks were developed which allow for an appropriate interpretation of students' understanding of the purpose of models.

Keywords: Models, Biology education, Students' understanding, Diagnosis, Forced Choice tasks, Validity

1 Einleitung

Die *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) nennen Diagnose als eine von vielen Einsatzmöglichkeiten von Messinstrumenten in den Erziehungswissenschaften. Bei einer Diagnose wird die Erfassung und Interpretation von Merkmalsausprägungen verknüpft mit der Generierung handlungsrelevanter Rückmeldungen (Fleischer, Koeppen, Kenk, Klieme & Leutner, 2013). Der Einsatz von Messinstrumenten zur Diagnose ist traditionell jedoch mit wenigen Ausnahmen dem Bereich der Psychologie vorbehalten (Gorin, 2007). In den Erziehungswissenschaften dagegen werden Messinstrumente bislang häufig genutzt, um unter dem Paradigma der Outcomeorientierung Kompetenzen in large-scale Studien zu erfassen und zu beschreiben. Entsprechend wird in vielen Lehrbüchern der Erziehungs- und Sozialwissenschaften die Entwicklung von large-scale Messinstrumenten beschrieben (Gorin, 2007).

In den letzten Jahren stieg die Nachfrage nach einer Anwendung der vorhandenen, theoretisch beschriebenen Kompetenzmodelle für die schulische Praxis (Hartig, Klieme & Leutner, 2008). Entsprechend sollen Kompetenzausprägungen von Schüler_innen für die unterrichtliche Praxis nutzbar gemacht werden. Als Alternative zu large-scale Untersuchungen, die aufgrund der durch Multi-Matrix-Designs bedingten Messfehler nicht für Individualdiagnosen verwendet werden können (Kauertz, Neumann & Haertig, 2012), gilt es nun, effiziente Diagnoseinstrumente zu entwickeln, die individuelle Kompetenzausprägungen erfassen und differenzierte Ansatzpunkte zur Förderung liefern (Fleischer et al., 2013; Hartig et al., 2008). Unter methodischer Perspektive ergibt sich hieraus ein Bedarf an Prozeduren, die die Entwicklung aussagekräftiger Diagnoseinstrumente gewährleisten (Gorin, 2007; Leighton & Gierl, 2007). Dabei wird gefordert, die Schüler_innen stärker in den Prozess der Aufgabenentwicklung einzubeziehen (Adams & Wieman, 2011; Leighton & Gierl, 2007). Der National Research Council (NRC, 2001) macht das Übertragungspotential dieses Vorgehens auf Bildungskontexte in seinem Standardwerk zur Testentwicklung deutlich: „The methods used in cognitive science to design tasks […] are applicable to many of the challenges of designing effective educational assessments." (NRC, 2001, S. 5).

Dieser Beitrag stellt am Beispiel von Forced Choice Aufgaben zur Diagnose einer Teilkompetenz der Modellkompetenz, dem Zweck von Modellen, eine mehrschrittige Prozedur vor, die eine empirische Überprüfung der Aufgaben durch Schüler_innen mit einschließt. Diese Prozedur beruht auf den *Standards for Educational and Psychological Testing* (AERA et al., 2014) und erfüllt die vom NRC (2001) geforderten Voraussetzungen für die Entwicklung von Messinstrumenten (vgl. 2.2). In einem ersten Schritt werden ausgehend vom Kompetenzmodell der Modellkompetenz (Upmeier zu Belzen & Krüger, 2010) Antwortalternativen zum Zweck von verschiedenen biologischen Modellen konstruiert, die auf verschiedene Niveaus von Modellkompetenz und – innerhalb der Niveaus – auf verschiedene inhaltliche Aspekte des Modells fokussieren. Zur anschließenden Überprüfung der Antwortalternativen wird mittels kurzer Interviews untersucht, ob die Schüler_innen die Formulierungen wie intendiert verstehen. Auf der Basis von Ratingaufgaben wird überprüft, inwiefern die Antwortalternativen für Schüler_innen relevante inhaltliche Aspekte des jeweiligen Modells repräsentieren. Dieses Vorgehen dient der Beurteilung von Validität und

wurde als Evidenzquelle genutzt, einerseits Antwortprozesse und andererseits den Testinhalt zu untersuchen (AERA et al. 2014). Abschließend werden geeignete Antwortalternativen in Forced Choice Aufgaben zusammengestellt.

2 Theoretischer Hintergrund

2.1 Zweck von Modellen

In der Wissenschaft werden Modelle als Hilfsmittel zur wissenschaftlichen Kommunikation und als Forschungsinstrumente genutzt (u. a. Harrison & Treagust, 2000; Passmore, Gouvea & Giere, 2014). Die Gründe, Modelle auch in der Schule zu nutzen, sind vielfältig. Modelle können zum einen bekannte Sachverhalte darstellen und sind damit Modelle *von* etwas (Mahr, 2008; Passmore, Gouvea & Giere, 2014). Damit eignen sie sich, um naturwissenschaftliches Wissen zu lernen (*learn science*; Hodson, 2014). Zum anderen können Modelle als Modelle *für* etwas gesehen werden (Mahr, 2008; Passmore, Gouvea & Giere, 2014), da aus Modellen Hypothesen abgeleitet werden, die mit Hilfe neuer Beobachtungen und Untersuchungen zur Konstruktion des erweiterten oder erneuerten Wissens führen (*learn to do science*; Hodson, 2014). Durch die Reflexion über die wissenschaftliche Arbeitsweise mit Modellen können Schüler_innen etwas über das Vorgehen in der Naturwissenschaft lernen (*learn about science*; Hodson, 2014).

Empirische Studien mit Schüler_innen zeigen, dass die Rolle von Modellen im wissenschaftlichen Erkenntnisprozess kaum wahrgenommen wird (u. a. Grosslight, Jay, Unger & Smith, 1991; Grünkorn, 2014; Treagust, Chittleborough & Mamiala, 2002; Trier & Upmeier zu Belzen, 2009). Grünkorn (2014) erfasste mittels offener Aufgaben Perspektiven, die Schüler_innen ($N = 706$) zum Zweck biologischer Modelle äußern. Hierbei gaben die befragten Schüler_innen an, der Zweck von Modellen sei entweder die Darstellung eines Sachverhalts oder das Erklären von Zusammenhängen. Weniger Schüler_innen äußerten, dass Modelle zum Überprüfen von Ideen genutzt werden könnten. Studien mit Lehrkräften naturwissenschaftlicher Fächer ergaben ebenfalls, dass Modelle im Unterricht vor allem als Medien zur Veranschaulichung eingesetzt werden (u. a. Crawford & Cullin, 2005; Justi & Gilbert, 2005; van Driel & Verloop, 2002). Trotzdem setzen Lehrkräfte ihrer eigenen Einschätzung nach Modelle im Unterricht auch wissenschaftlich für die Erkenntnisgewinnung ein (Krell & Krüger, 2013).

Nach dem Kompetenzmodell der Modellkompetenz (Upmeier zu Belzen & Krüger, 2010) werden fünf Teilkompetenzen (Eigenschaften von Modellen, Alternative Modelle, Zweck von Modellen, Testen von Modellen, Ändern von Modellen) unterschieden, die in drei Niveaus unterschiedliche Perspektiven auf Modelle beschreiben. In der Teilkompetenz „Zweck von Modellen“ (Tab. 1) bilden sich die zuvor genannten Schülerperspektiven ab. Niveaus I und II beschreiben einen vom Modellierer oder Modellnutzer definierten Anspruch an das Modell als Modell *von* etwas und in Niveau III den Anspruch an ein Modell als Modell *für* etwas (Upmeier zu Belzen & Krüger, 2010). Die Bildungsstandards (KMK, 2005) fordern explizit beide Perspektiven auf Modelle. Eine elaborierte Modellkompetenz zeigt sich dabei darin, dass Schüler_innen mit Modellen auch hypothesenbasiert im Sinne des Niveaus III argumentieren können.

Tabelle 1: Niveaus der Teilkompetenz „Zweck von Modellen“ (Upmeier zu Belzen & Krüger, 2010).

Niveau I	Niveau II	Niveau III
Modellobjekt zur Beschreibung von etwas einsetzen	Bekannte Zusammenhänge und Korrelationen von Variablen im Ausgangsobjekt erklären	Zusammenhänge von Variablen für zukünftige neue Erkenntnisse voraussagen

2.2 Entwicklung und Überprüfung von Diagnoseaufgaben

Eine individuelle Förderung von Schüler_innen bzgl. ihres Modellverstehens setzt dessen Diagnose voraus. Dabei sollen einerseits Vorstellungen zum Themenbereich ermittelt werden, die für Lehr- und Lernprozesse relevant sein können, und andererseits sollen Lernergebnisse festgestellt werden, die sich nach Vermittlungssituationen ergeben. Die Diagnose vor dem Lernprozess kann je nach Vorstellungs- oder Fähigkeitsprofil zu Zuweisungen der Schüler_innen zu Lerngruppen führen. In diesen Lerngruppen lässt sich individuelles Lernen mit angepassten Fördermaßnahmen optimieren (vgl. Ingenkamp & Lissmann, 2008).

Eine theoretische Basis für die Diagnose und Förderung der Fähigkeiten von Schüler_innen bilden Kompetenzmodelle. Diese beschreiben „Zusammenhänge zwischen individuellen Fähigkeiten und Fertigkeiten und erfolgreichem Handeln in spezifischen Kontexten“ (Klieme & Hartig, 2007, S. 11). Explizite, handlungsrelevante Rückmeldungen an Lehrkräfte, wie beispielsweise die Zuweisung der Schüler_innen zu spezifischen Fördermaßnahmen, lassen sich aber nicht direkt aus den Modellen ableiten (vgl. Kauertz et al., 2012). Für diese Art der Rückmeldungen müssen individuelle Kompetenzausprägungen empirisch durch Diagnoseinstrumente bestimmt werden. Adams & Wieman (2011) betonen, dass Diagnoseinstrumente, die einen unterrichtlich nutzbaren Informationsgewinn schaffen, effizient ohne Training einsetzbar sowie valide interpretierbar sein müssen. Sie empfehlen hierfür den Einsatz von Aufgaben mit geschlossenem Antwortformat, da diese einfacher zu bearbeiten und auszuwerten sind. Für eine valide Interpretation der Diagnoseergebnisse ist bereits während der Aufgabenentwicklung sicherzustellen, dass (1) die Schüler_innen die Fähigkeit, die mit dem Verfahren gemessen werden soll, auch tatsächlich bei der Bearbeitung der Aufgaben einsetzen (Antwortprozesse; AERA et al., 2014) und (2) die Aufgaben das zu messende Konstrukt adäquat abbilden (Testinhalt; AERA et al., 2014).

Da Validität die Interpretation der Ergebnisse des Diagnoseinstruments betrifft, sollte das Verständnis der Diagnoseaufgaben empirisch mit der Stichprobe getestet werden, für die das Instrument konzipiert wurde (Antwortprozesse; AERA et al., 2014). Adams & Wieman (2011) schlagen vor, zum Zweck der Validierung während der Aufgabenentwicklung Schülerinterviews durchzuführen, in denen Schüler_innen ihre Antworten begründen oder erklären, wie sie bestimmte Begriffe verstehen. Sie verweisen in diesem Zusammenhang auch auf Ericsson und Simon (1998) und betonen: „Because these sorts of probing questions do alter student thinking and could likely help students think of connections they may not have in

an actual testing situation, strict think-aloud interviews must be performed for validation once the test is constructed." (Adams & Wieman, 2011, S. 1297).

Um darüber hinaus Hinweise für Validität auf der Basis des Testinhalts zu erhalten, muss überprüft werden, ob die Aufgaben das zu messende Konstrukt adäquat abbilden (Testinhalt; AERA et al., 2014). Bei der Auswahl und Überprüfung von Aufgaben, die sich auf bestimmte Kontexte beziehen, sollte berücksichtigt werden, dass unterschiedliche Kontexte jeweils spezifische kognitive Anforderungen transportieren und dies einen Einfluss auf die Perspektiven von Schüler_innen haben kann (Krell, Upmeier zu Belzen & Krüger, 2014; Nehm & Ha, 2011). In der vorliegenden Studie wird die Kontextualisierung der Aufgaben durch die Verwendung unterschiedlicher biologischer Modelle in den Aufgabenstämmen realisiert. Der Begriff Kontext wird hier folglich im Sinne eines Item-Features genutzt (Krell, Upmeier zu Belzen & Krüger, 2012). Bei Aufgaben mit geschlossenem Antwortformat empfiehlt es sich, Antwortalternativen zu konstruieren, die die Perspektiven von Schüler_innen berücksichtigen (Haladyna, 2004). Hierbei können auf der einen Seite die Antwortalternativen auf der Grundlage zuvor erhobener Schüleraussagen konstruiert (Haladyna, 2004) und anschließend durch Experten in Bezug auf die Passung mit der zugrundeliegenden Theorie überprüft werden. Auf der anderen Seite kann nachträglich überprüft werden, ob die konstruierten und von Experten überprüften Antwortalternativen von den Schüler_innen tatsächlich als relevante Perspektiven in Bezug auf den Kontext der Aufgabe beurteilt werden (Leighton & Gierl, 2007).

3 Forschungsfragen und Hypothesen

Für eine effiziente und zugleich individuelle Diagnose von Modellverstehen sollen Aufgaben im Forced Choice Format eingesetzt werden, die je eine Antwortalternative pro Niveau zum Zweck von Modellen enthalten. Schüler_innen sollen aus den drei Antwortalternativen diejenige auswählen, die ihrer eigenen Meinung am ehesten entspricht (McCloy, Heggestad & Reeve, 2005). Für die Entwicklung solcher Diagnoseaufgaben wird zunächst ein Pool von Antwortalternativen für verschiedene biologische Kontexte konstruiert. Hierbei entstehen pro Niveau und Kontext mehrere Antwortalternativen, die auf verschiedene inhaltliche Aspekte des Kontexts fokussieren. Um die Antwortalternativen zu evaluieren, zu selektieren und eine Auswahl in Forced Choice Aufgaben zusammenzustellen, sind zwei Forschungsfragen empirisch zu untersuchen.

F1: Inwiefern verstehen die Schüler_innen die konstruierten Antwortalternativen auf dem jeweils theoretisch intendierten Niveau?

F2: Inwieweit unterscheiden sich die konstruierten Antwortalternativen in ihrer Relevanz für Schüler_innen in Bezug auf den jeweiligen biologischen Kontext?

Es wird erwartet, dass die Schüler_innen die einzelnen Antwortalternativen entsprechend des theoretisch intendierten Niveaus verstehen, womit eine auf den Antwortprozessen basierende Evidenz für Validität erbracht wäre (AERA et al., 2014; Leighton & Gierl, 2007).
Um das Modellverstehen von Schüler_innen erfassen zu können, sollten – unabhängig vom Niveau – die in den Antwortalternativen fokussierten inhaltlichen Aspekte für Schüler_innen in Bezug auf den jeweiligen biologischen Kontext möglichst relevant sein. In Anlehnung an Studien, die inhaltliche Aspekte als schwierigkeitserzeugendes Aufgabenmerkmal betrachten, wird vermutet, dass der in den verschiedenen Antwortalternativen enthaltene Inhalt einen Einfluss darauf hat, wie relevant Schüler_innen die jeweilige Antwortalternative wahrnehmen (Cohors-Fresenborg, Sjuts & Sommer, 2004; Kauertz, 2008; Krell et al., 2014).

4 Methodisches Vorgehen

4.1 Entwicklung von Antwortalternativen

Als Ausgangspunkt für die Entwicklung der Forced Choice Aufgaben wurden für acht verschiedene biologische Kontexte je sechs Antwortalternativen konstruiert. Dabei wurden für jedes der drei Niveaus der Teilkompetenz „Zweck von Modellen“ (Tab. 1; Upmeier zu Belzen & Krüger, 2010) je zwei niveaugleiche Formulierungen entwickelt. Diese unterscheiden sich zum einen im niveaubestimmenden Verb, z. B. „sichtbar machen“ und „veranschaulichen“ als Indikatoren für Niveau I. Zum anderen beziehen sich die Antwortalternativen auf verschiedene Aspekte des in der Aufgabe genutzten Modells, z. B. auf den „Aufbau“ oder auf die „Bestandteile“. Nach ihrer empirischen Überprüfung sollte je eine Antwortalternative pro Niveau in eine Forced Choice Aufgabe implementiert werden (Tab. 2).
Es wurden Kontexte gewählt, die eine sinnvolle Formulierung von Antwortalternativen auf allen drei Niveaus der Teilkompetenz „Zweck von Modellen“ gemäß des Kompetenzmodells der Modellkompetenz erlauben.

Tabelle 2: Beispielhafte Darstellung der sechs entwickelten Antwortalternativen für den Kontext der Biomembran. Die niveaubestimmenden Verben sind fett gedruckt.

Das Modell der Biomembran hat den Zweck …	
Niveau I	… *den Aufbau der Biomembran* ***sichtbar zu machen.*** [Ia] … *die verschiedenen Bestandteile der Biomembran* ***zu veranschaulichen****.* [Ib]
Niveau II	… *den Aufbau der Biomembran* ***zu erklären****.* [IIa] … *das Zusammenwirken der Bestandteile* ***begreiflich zu machen****.* [IIb]
Niveau III	… *den Aufbau der Biomembran* ***weiter zu erforschen.*** [IIIa] … *weitere Bestandteile der Biomembran* ***vorauszusagen.*** [IIIb]

4.2 Datenerhebung

Für die Untersuchung der Forschungsfragen und die empirische Überprüfung der Antwortalternativen unter Berücksichtigung der Einschätzungen von Schüler_innen wurden ein quantitativer und ein qualitativer Ansatz kombiniert (*convergent mixed method design*; Creswell & Plano Clark, 2011).

Für den quantitativen Ansatz wurden die jeweils sechs konstruierten Antwortalternativen pro Kontext in zufälliger Reihenfolge in Ratingaufgaben zusammengestellt (Abb. 1). Jede Aufgabe bestand aus einem standardisierten Aufgabenstamm, einer Abbildung des Modells sowie des Originals, einem standardisierten Impuls für die Aufgabenbearbeitung und den sechs Antwortalternativen.

In der linken Abbildung siehst du eine mikroskopische Aufnahme einer Biomembran und in der rechten Abbildung ein Modell der Biomembran, das Biologen entworfen haben.					
Abbildung 1: Mikroskopische Aufnahme einer Biomembran Abbildung 2: Modell der Biomembran					
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, <u>welchen Zweck</u> dieses Modell der Biomembran haben kann!					
Entscheide bei jeder Aussage, wie sehr sie deiner eigenen Meinung entspricht. *Mache neben jeder Aussage nur ein Kreuz!*					
Das Modell der Biomembran hat den Zweck ...	*gar nicht*	*wenig*	*teils-teils*	*annähernd*	*völlig*
… den Aufbau der Biomembran sichtbar zu machen. [Ia]	□	□	□	□	□
… die verschiedenen Bestandteile der Biomembran zu veranschaulichen. [Ib]	□	□	□	□	□
… den Aufbau der Biomembran zu erklären. [IIa]	□	□	□	□	□
… das Zusammenwirken der Bestandteile begreiflich zu machen. [IIb]	□	□	□	□	□
… den Aufbau der Biomembran weiter zu erforschen. [IIIa]	□	□	□	□	□
… weitere Bestandteile der Biomembran vorauszusagen. [IIIb]	□	□	□	□	□

Abbildung 1: Aufgabenbeispiel zum Kontext der Biomembran. Die Angabe der Antwortalternativen (Ia, Ib, IIa, …) dient der Illustration. Die Reihenfolge der Antwortalternativen entspricht nicht der Reihenfolge im Fragebogen.

Insgesamt bewerteten $N = 275$ Schüler_innen der Klassenstufen neun bis zwölf an Berliner Gymnasien auf einer fünfstufigen, verbaläquidistanten Likert-Skala (Abb. 1; vgl. Moosbrugger & Kelava, 2012), wie sehr die einzelnen Antwortalternativen ihrer eigenen Meinung entsprechen. Um die Arbeitsbelastung der Schüler_innen zu reduzieren, wurden die acht Aufgaben auf vier Testhefte aufgeteilt, die jeweils vier Aufgaben enthielten (*balanced incomplete block design*; Gonzalez & Rutkowski, 2010).

Für den qualitativen Ansatz wurden im Anschluss an die schriftliche Befragung mit je $n = 40$ Schüler_innen pro Klassenstufe vollstrukturierte Interviews geführt. Hierfür unterstützte der eigene, zuvor ausgefüllte Fragebogen als *Stimulated Recall* die Erinnerung an das Bearbeiten der Aufgaben (Sandmann, 2014). Im Interview lasen die Schüler_innen zunächst zur ersten Aufgabe die erste Antwortalternative und dann ihre Bewertung dieser auf der Likert-Skala vor. Anschließend wurden die Schüler_innen gebeten, ihre Bewertung zu begründen. Das gleiche Vorgehen wurde für alle sechs Antwortalternativen für zwei der vier im Fragebogen bearbeiteten Aufgaben durchgeführt. Insgesamt ergeben sich bei $n = 160$ interviewten Schüler_innen 1920 Schüleraussagen, folglich 40 Aussagen pro Antwortalternative.

4.3 Datenauswertung

Die Datenauswertung und die damit einhergehende sukzessive Selektion der Antwortalternativen verliefen zweigeteilt.

Um zu untersuchen, ob die Schüler_innen die einzelnen Antwortalternativen entsprechend des theoretisch intendierten Niveaus verstehen (Forschungsfrage 1), wurden die Audiodaten transkribiert und die Schüleraussagen mittels eines Leitfadens binär in „entsprechend der Theorie verstanden (1)“ oder „nicht entsprechend der Theorie verstanden (0)“ kodiert. Es war nicht von Interesse, ob das Niveau der Antwortalternative dem Niveau des Modellverstehens der Schüler_in entsprach. War aus der Schüleraussage nicht zu entnehmen, ob die Antwortalternative im Sinne des Niveaus verstanden wurde, wurde mit „8“ kodiert. 50 % der Aussagen wurden durch einen unabhängigen Rater zweitkodiert. Die Übereinstimmung der Rater wurde mittels Cohens Kappa überprüft (Wirtz & Caspar, 2002). Anschließend wurde ermittelt, wie viele Schüler_innen eine Antwortalternative entsprechend der Theorie verstanden haben. Es wurde zuvor in einer Expertenrunde (2 Wissenschaftler_innen aus dem Bereich der Biologiedidaktik, 2 Wissenschaftler_innen aus der Psychologie) festgelegt, diejenigen Antwortalternativen für eine weitere Verwendung in der Diagnose als geeignet einzustufen, die von mindestens 70 % der Schüler_innen entsprechend der Theorie verstanden wurden. Alle Antwortalternativen, die dieser Voraussetzung nicht genügten, wurden verworfen.

Verstanden die Schüler_innen beide Antwortalternativen eines Niveaus hinreichend gut, folgte ein weiterer Selektionsschritt, der die Relevanz der Antwortalternativen für die Schüler_innen (Forschungsfrage 2) nutzte. Hierfür wurden die Zustimmungen der Schüler_innen zu den zwei niveaugleichen Antwortalternativen auf der Likert-Skala (1-5; ca. 140 Bewertungen pro Antwortalternative) verglichen und damit auf die Relevanz der Antwortalternativen in Bezug auf den Kontext geschlossen. Für die Zusammenstellung der Forced Choice Aufgaben wurden diejenigen Antwortalternativen ausgewählt, die im

Vergleich als relevanter in Bezug auf den angebotenen Kontext eingestuft wurden und somit am ehesten (vgl. McCloy et al., 2005) den Perspektiven von Schüler_innen beim jeweiligen Kontext entsprachen.

5 Ergebnisse

5.1 Passung zwischen intendiertem und interpretiertem Niveau

Die qualitative Bewertung der Schüleraussagen durch zwei unabhängige Rater wurde mit einer sehr guten Interrater-Reliabilität (Cohens-Kappa $0{,}86 < \kappa < 0{,}93$; Wirtz & Caspar, 2002) durchgeführt. Um die genutzten Kategorien vorzustellen, werden aus den Schüleraussagen ($N = 1920$) beispielhaft einige für die Kodierungen zu einer der beiden Antwortalternativen auf Niveau III für den Kontext der Biomembran vorgestellt (Tab. 3).

Tabelle 3: Beispielhafte Kodierung von Schüleraussagen zu einer Antwortalternative auf Niveau III beim Kontext der Biomembran.

Das Modell der Biomembran hat den Zweck, weitere Bestandteile der Biomembran vorauszusagen. [Antwortalternative IIIb]	
Kodierung	Ankerbeispiel
1 - Antwortalternative wurde entsprechend der theoretischen Intention verstanden	**Maon:** Weil ich dachte, da geht es darum, dass man jetzt sagen kann, in einem bestimmten Abstand kommt jetzt bestimmt wieder so ein blaues Dings-Bums da. **Bean:** Ich denke, das Modell dient nur dazu, das Ganze zu veranschaulichen, aber man kann anhand eines Modells keine Vorhersagen treffen.
0 - Antwortalternative wurde nicht entsprechend der theoretischen Intention verstanden	**Kaix:** Man erkennt im Modell zwar weitere [Bestandteile], die sind aber nicht beschriftet oder irgendwas. Und vorauszusagen war auch so ein bisschen komisch formuliert. - *Was findest du daran komisch?* - Naja vorauszusagen. Wenn, dann hätte ich jetzt anzugeben oder sowas eingesetzt, weil das ist irgendwie ein bisschen klarer dann.
8 - Antwort nicht aussagekräftig	**Kasa:** Ich habe mir davor noch die anderen [Antwortalternativen] durchgelesen und dachte, da passt eher so der Aufbau.

In der Aussage von Bean wird deutlich, dass es möglich ist, eine Antwortalternative auf dem theoretisch intendierten Niveau III zu verstehen, ohne selbst ein Modellverstehen auf Niveau III zu besitzen. Gleiches geschah bei Niveau II, wenn Schüler_innen anerkannten, dass man mit dem Modell auch erklären könnte, den Zweck dieses speziellen Modells aber nicht in einer Erklärung, sondern in der Beschreibung des Originals sahen. So sagt Gura: *Das erklärt nichts. Da ist keine Schrift, nichts beschrieben. Ich habe da eine Abbildung*.
Schüler_innen, die die Antwortalternative nicht entsprechend des Niveaus III verstanden, argumentierten hauptsächlich mit Bezug auf das Verb „voraussagen“, ohne dabei den hypothetischen Charakter der Aussage, welcher durch eben jenes Verb beschrieben werden

sollte, wiederzugeben. Sie interpretierten die Bedeutung des Verbs in eine Bedeutung um, welches den von ihnen gedachten Zweck des Modells besser beschreibt; in der Aussage von Kaix in das Verb „angeben“ (Tab. 3).
Auch bei den angebotenen Antwortalternativen auf Niveau II, die nicht entsprechend der Theorie verstanden wurden, erkannten die Schüler_innen den Bedeutungsunterschied zwischen den Verben „zeigen/ darstellen/ wiedergeben“ und „erklären/ erläutern/ verständlich machen“ nicht und nutzen die Verben synonym. So begründete Sore seine völlige Zustimmung zur Niveau II Aussage „Das Modell der Biomembran hat den Zweck, den Aufbau der Biomembran zu erklären“ wie folgt: *Die Bestandteile werden alle gezeigt und daran kann man veranschaulichen, wo die ganzen Proteine sich in der Membran befinden. Das kann man gut beschreiben.* Sore interpretierte die Antwortalternative in Niveau I um.
Zum Zweck der Selektion wurde für jede Antwortalternative eines jeden Kontexts ermittelt, wieviel Prozent der jeweils 40 interviewten Schüler_innen diese entsprechend des theoretisch intendierten Niveaus verstanden haben (Tab. 4).

Tabelle 4: Häufigkeit [%], mit der eine Antwortalternative entsprechend des theoretisch intendierten Niveaus interpretiert wurde. Graue Unterlegung: mehr als 70 %. Klammer: Anzahl der nicht interpretierbaren Aussagen von 40. $N_{Aussagen} = 1920$.

Kontext		Niveau I		Niveau II		Niveau III	
		Ia	Ib	IIa	IIb	IIIa	IIIb
Biomembran	[BM]	**98** (1)	93 (2)	**83** (2)	90 (4)	**75** (5)	53 (6)
Evolution	[EV]	75 (10)	**83** (4)	**74** (9)	85 (6)	24 (17)	**73** (7)
Gehirn	[GH]	**100** (0)	78 (9)	**73** (9)	44 (19)	**79** (4)	33 (9)
Jurawald	[JW]	**90** (4)	88 (5)	**70** (9)	53 (18)	38 (13)	**72** (7)
T. rex	[TR]	83 (7)	**83** (7)	69 (9)	**75** (10)	35 (13)	**70** (9)
Uferzone	[UF]	90 (4)	**95** (2)	**70** (10)	67 (12)	3 (9)	**70** (10)
Luftstrom	[LS]	63 (15)	58 (16)	53 (19)	53 (19)	63 (11)	45 (16)
Bakterienwachstum	[BW]	58 (14)	58 (16)	45 (16)	43 (18)	58 (13)	65 (12)

Legt man den für die Selektion definierten Richtwert an, dass mindestens 70 % der Schüler_innen eine Antwortalternative entsprechend des intendierten Niveaus verstanden haben müssen, dann findet man mit Ausnahme der Kontexte „Luftstrom“ und „Bakterienwachstum“ für jeden Kontext pro Niveau mindestens eine Antwortalternative, bei deren späterem Einsatz in Forced Choice Aufgaben ein zufriedenstellendes Verständnis der Schüler_innen erwartet werden kann.
Bei den Kontexten „Luftstrom“ und „Bakterienwachstum“ handelt es sich um Modelle, die als Formeln präsentiert werden. Beim „Luftstrom“ wurde beispielsweise ein Schwimmer als Original dargestellt und eine Formel zur Menge der Luft in der Lunge beim Atmen. Viele Schüleraussagen waren für die Auswertung nicht aussagekräftig (Kodierung „8“; siehe Tab. 4), da die Schüler_innen die Formeln nicht als Modell verstanden. Der Proband Jaul sagt

beispielsweise zu seiner Entscheidung in der Ratingaufgabe zum Kontext „Luftstrom“: *Da würde ich dann doch eher für das Modell hier [zeigt auf das Original] stimmen, weil da sieht man einen Menschen, der gerade ausatmet. Das [zeigt auf das Modell] ist hier eher wie eine Formel und ich finde da kann ich nicht so viel rauslesen, ehrlich gesagt, aus dieser Formel.* Alle Antwortalternativen zu den beiden Kontexten wurden im ersten Selektionsschritt verworfen.

Unterschiede zwischen theoretischer Intention und Interpretation durch die Schüler_innen zeigen sich sowohl zwischen den Niveaus als auch zwischen einzelnen niveaugleichen Antwortalternativen innerhalb eines Kontexts. Bei den sechs verbleibenden Kontexten nimmt bei höherem Niveau das theoriekonforme Verständnis der Antwortalternativen grundsätzlich ab. Auf Niveau III wird bei allen Kontexten jeweils nur eine der beiden Antwortalternativen von mind. 70 % der Schüler_innen entsprechend des Niveaus III interpretiert.

Im nächsten Schritt wurde pro Niveau die Relevanz der verstandenen Antwortalternativen untersucht. Für die Kontexte „Gehirn“, „Jurawald“, „T. rex“ und „Uferzone“ geschah dies ausschließlich auf Niveau I, für die Kontexte „Biomembran“ und „Evolution“ auf den Niveaus I und II.

5.2 Relevanz niveaugleicher Antwortalternativen

Vergleicht man die Bewertungen der Schüler_innen für die nach dem ersten Schritt verbleibenden Antwortalternativen pro Niveau pro Kontext auf der Likert-Skala, zeigen sich zum Teil deutliche Unterschiede (Abb. 2).

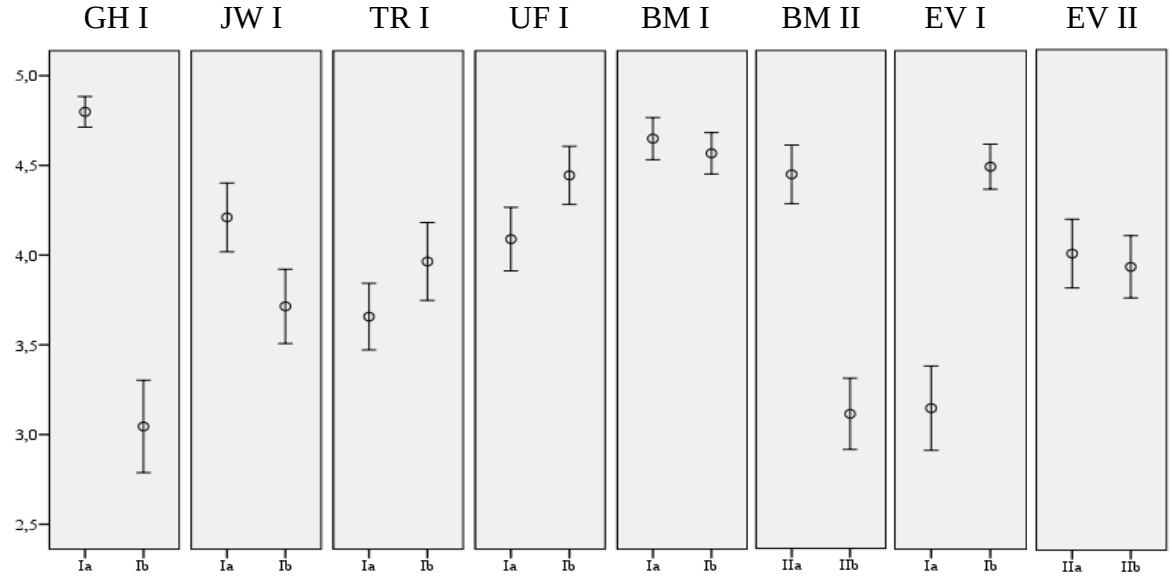

Abbildung 2: Bewertung der Antwortalternativen a und b pro Niveau [I, II, III] und pro Kontext (Abkürzungen [GH – EV] siehe Tab. 4.) auf der Likert-Skala (1: gar nicht, 2: wenig, 3: teils-teils, 4: annähernd, 5: völlig).

Beim Kontext der Biomembran (BM) wurde der Antwortalternative IIa auf der Likert-Skala deutlich mehr zugestimmt als der Antwortalternative IIb. Die Schüler_innen geben folglich an, dass das Modell der Biomembran eher den Zweck habe, den Aufbau der Biomembran zu erklären (Antwortalternative IIa), als das Zusammenwirken der Bestandteile begreiflich zu

machen (Antwortalternative IIb). In den mündlichen Begründungen argumentieren die Schüler_innen zwar, dass mit Hilfe des Modells auch einzelne Bestandteile erklärt werden können, das Zusammenwirken dieser mache das Modell aber durch das Fehlen schriftlicher Erklärungen oder der für das Verständnis von Prozessen notwendigen Animationen nicht begreiflich. Der Schüler Amon sagt im Interview dazu: *Das ist ein Bild und darauf bewegt sich nichts und dazu steht hier auch nichts. Demnach könnte ich mir so nicht erklären, wie die [Bestandteile] zusammenwirken*. Es wird deutlich, dass die Antwortalternative IIb von Amon entsprechend des theoretisch intendierten Niveaus II verstanden wurde, für ihn jedoch Mängel an diesem speziellen Modell den angegebenen Zweck für dieses Modell unterbinden. Beim Kontext Evolution (EV) stimmten die Schüler_innen den beiden Antwortalternativen auf Niveau II in etwa gleich stark zu. Antwortalternative IIa wurde selektiert, da sie sich auf den gleichen Aspekt des Kontexts bezieht, wie die Antwortalternativen aus den anderen Niveaus.
Für alle Forced Choice Aufgaben wurde aus den in Abbildung 2 dargestellten Antwortalternativen jeweils die auf der Likert-Skala höher bewertete und damit relevantere pro Kontext und Niveau ausgewählt. Die ausgewählten Antwortalternativen pro Kontext auf den Niveaus I-III sind in Tabelle 4 fett markiert.

6 Diskussion

Die Untersuchung beider Forschungsfragen trug dazu bei, insgesamt sechs Forced Choice Aufgaben zu verschiedenen biologischen Kontexten zu entwickeln. Die vorgestellte Prozedur schloss eine empirische Überprüfung der Eignung der Aufgaben für die Diagnose des Modellverstehens in der Teilkompetenz „Zweck von Modellen“ durch Schüler_innen mit ein. Durch die Kombination eines quantitativen (Ratingskala) und eines qualitativen Ansatzes (Schülerinterviews) in einem *convergent mixed method design* (Creswell & Plano Clark, 2011) wurde eine Vielzahl von Schüleraussagen (N=1920) bei der Aufgabenentwicklung berücksichtigt. Dies kommt der Forderung nach, die Schüler_innen stärker in den Prozess der Aufgabenentwicklung einzubeziehen (Adams & Wieman, 2011; Leighton & Gierl, 2007).
Die Schüler_innen erklärten in Interviews, wie sie die einzelnen Antwortalternativen verstehen und lieferten dadurch bereits während der Aufgabenentwicklung Hinweise auf Validität. Die Ergebnisse dieses Vorgehens können jedoch nicht mit Protokollen lauten Denkens gleichgesetzt werden. Ericsson und Simon (1998) unterscheiden in Anlehnung an Vygotsky (1962) zwischen *inner speech* und *social speech* und betonen, dass die Verbalisierung der eigenen Gedankenprozesse in einer Gesprächssituation wie z. B. einem Interview (*social speech*) die Gedanken selbst ändern bzw. beeinflussen können. Kritiker schlussfolgern, dass es nicht möglich sei, die aufeinander folgenden Gedanken der Schüler_innen bei der Aufgabenbearbeitung zu erfassen, die zur Lösung führen (Ericsson & Simon, 1998). Diese Kritik kann für die vorliegende Studie dahingehend ausgeklammert werden, dass bei der Erhebung der Schüleraussagen nicht beabsichtigt wurde, den komplexen Gedankenprozess bei der Aufgabenbearbeitung nachzuvollziehen. Vielmehr sollte hier geprüft werden, ob die konstruierten Antwortalternativen wie intendiert verstanden wurden. Hierfür liefern die Interviewdaten interpretierbare und aussagekräftige Ansatzpunkte (Adams & Wieman, 2011).

Bezogen auf die Ergebnisse der Untersuchung sind vor allem drei inhaltliche Aspekte zu diskutieren. Zum einen nimmt die theoriekonforme Interpretation der Antwortalternativen mit steigendem Niveau ab. Zum anderen scheinen die angebotenen Kontexte unterschiedliche kognitive Anforderungen zu transportieren, was teilweise zu Verständnisproblemen führt (Krell et al., 2014; Nehm & Ha, 2011). Es muss diskutiert werden, warum niveaugleiche Antwortalternativen zum gleichen Kontext für die Schüler_innen unterschiedlich relevant sind.

6.1 Einfluss des Niveaus auf die Interpretation der Antwortalternativen

Insgesamt zeigt sich eine gute Übereinstimmung zwischen dem theoretisch intendierten Niveau in den Antwortalternativen und dem Verstehen dieser durch die Schüler_innen (Forschungsfrage 1). Folglich können viele der konstruierten Antwortalternativen dazu dienen, Rückschlüsse auf das Verstehen der Schüler_innen zu ziehen (Tab. 4; AERA et al., 2014). Nichtsdestotrotz nimmt die Passung des interpretierten mit dem theoretisch intendierten Niveau der Antwortalternativen mit höheren Niveaus ab. Als ein Erklärungsansatz könnte ein geringes Modellverstehen von Seiten der Schüler_innen herangezogen werden. Geht man davon aus, dass Schüler_innen die Rolle von Modellen im wissenschaftlichen Erkenntnisprozess in der Schule kaum erleben (Crawford & Cullin, 2005; Justi & Gilbert, 2005; van Driel & Verloop, 2002) und Modelle daher nicht als Instrumente der Forschung ansehen (Grosslight et al., 1991; Grünkorn, 2014; Treagust et al., 2002; Trier & Upmeier zu Belzen, 2009), ist es möglich, dass sie dem Inhalt von Antwortalternativen, die die epistemologische Bedeutung von Modellen ausdrücken (Niveau III), nicht nur nicht zustimmen, sondern diesen Inhalt auch nicht entsprechend verstehen. Als Folge zeigt sich möglichweise, dass das niveauangebende Verb (z. B. vorhersagen, vermuten) von Schüler_innen ignoriert und die Zustimmung zu einer Antwortalternative allein nach deren Inhalt entschieden wird. Alternativ könnte eine Uminterpretation des niveauangebenden Verbs entsprechend des eigenen Modellverstehens erfolgen. Diese Art der Uminterpretation als Strategie wurde auch von Krell, Czeskleba und Krüger (2012) in einer Studie mit lautem Denken identifiziert. Die Forscher beobachteten, dass Schüler_innen in Paarvergleichs-aufgaben, sofern ihr präferiertes Niveau nicht angeboten wurde, eine andere Antwortalternative entsprechend des gewünschten Niveaus umdeuteten. Die valide Interpretation der Ergebnisse bezogen auf das Modellverstehen ist in diesen Fällen nicht möglich (AERA et al., 2014).

Im Hinblick auf die Nutzung der Forced Choice Aufgaben zur Diagnose von Modellverstehen führen Uminterpretationen oder Bewertungen auf der Basis des Inhalts der Antwortalternativen dazu, dass Schüler_innen ein nicht zutreffendes Modellverstehen zugeschrieben wird. Dies führt wiederum zu Problemen bei der adäquaten Förderung. Die Interviewdaten konnten hier genutzt werden, um Hinweise für die Auswahl von Antwortalternativen zu sammeln und damit diese Probleme zu reduzieren.

6.2 Einfluss des Kontexts auf die Interpretation der Antwortalternativen

Für die Kontexte „Luftstrom“ und „Bakterienwachstum“ war es nicht möglich, Antwortalternativen zu selektieren, die von den Schüler_innen in ausreichendem Maße wie intendiert verstanden wurden. Hier konnte die Hypothese, dass die Bearbeitung der Aufgaben später zu valide interpretierbaren Diagnoseergebnissen führt, nicht bestätigt werden (AERA et al., 2014). Gründe für das Nichtverstehen der Antwortalternativen scheinen nicht niveauspezifisch zu sein, da der Prozentsatz der Schüler_innen, die die Antwortalternativen nicht verstehen, sich zwischen den Niveaus I, II und III nicht unterscheidet. Es kann vermutet werden, dass die Repräsentationsform der Modelle (hier Formeln) ein Grund für die Verständnisprobleme war. Cohors-Fresenborg et al. (2004), die PISA Aufgaben im Hinblick auf kognitionsorientierte Aufgabenmerkmale untersuchten und klassifizierten, leiten aus ihren Ergebnissen ab, dass das Merkmal „Formalisierung von Wissen“ ein auf besondere Weise schwierigkeitserzeugendes Merkmal ist. Cohors-Fresenborg et al. (2004) betonen zudem, es handle sich bei formalisiertem Wissen auch um ein „Werkzeug, dessen Handhabung eine Kompetenz darstellt, Komplexität zu bewältigen“ (Cohors-Fresenborg et al., 2004, S. 121). Möglicherweise können die Schüler_innen mit diesem Werkzeug noch nicht entsprechend kompetent umgehen. Dies zeigt sich womöglich in der Schüleraussage von Jaul: *Weil das ja hier eher wie so eine Formel ist und ich finde, da kann ich nicht so viel rauslesen, ehrlich gesagt, aus dieser Formel.* Solche und ähnliche Aussagen lassen die Vermutung zu, dass die Aufgaben mit den formelhaften Modellen zu den Kontexten „Luftstrom“ und „Bakterienwachstum“ nicht nur das Modellverstehen der Schüler_innen, sondern, als eine Interaktion, möglicherweise auch deren Umgang mit Formeln messen. Durch die mangelnde Trennung der Konstrukte eignen sich diese Kontexte nicht zur Erfassung von Modellverstehen und wurden daher verworfen.

6.3 Unterschiede zwischen niveaugleichen Antwortalternativen zu einem Kontext

Scheinbar führen kontextspezifische inhaltliche Aspekte dazu, dass Schüler_innen eine von zwei niveaugleichen Antwortalternativen mehr oder weniger bevorzugen. Die Gründe könnten bei den inhaltlichen Eigenschaften der Antwortalternativen liegen. Die Inhalte einiger Antwortalternativen passen nach Meinung der Schüler_innen offensichtlich besser oder weniger gut zum dargestellten Modellobjekt als andere. Demnach wird bei einigen Antwortalternativen eine andere Repräsentation des Modells erwartet, z. B. eine Beschriftung oder Animation der Biomembran.

Außerdem spielen personenbezogene Merkmale wie Vorwissen und Interesse eine Rolle. Obwohl das Vorwissen bei dieser Studie nicht erhoben wurde, zeigen sich in den Interviewdaten einige interessante Schülervorstellungen, die sich möglicherweise auf die Bewertung der Antwortalternativen auswirken. Beim Kontext des Bakterienwachstums zeigt sich die Schülervorstellung, dass Lebewesen durch Vergrößerung der Zellen wachsen („Wachstum ist größer werden“; Riemeier, 2005) z. B. darin, dass viele Schüler_innen den Ausdruck „Menge an Bakterien“ kritisch bewerten. Douc begründet: *Da habe ich wenig angekreuzt, weil es um das Wachstum geht und nicht um die Menge an Bakterien.*

Mit Bezug zum eigenen Interesse kommentiert Jael die Bevorzugung der Antwortalternative „Das Modell des T. rex hat den Zweck, den Einfluss der Vorderbeine auf die Fortbewegungsart des T. rex verständlich zu machen.“ gegenüber der niveaugleichen Antwortalternative wie folgt: *[Es ist] ziemlich interessant, was die Vorderbeine eigentlich für einen Einfluss haben darauf, wie er sich bewegt.* Solche inhaltlichen Aspekte werden von Kauertz (2008) als potentiell schwierigkeitserzeugend beschrieben. Möglicherweise wirkt sich das Interesse nicht nur darauf aus, welche Antwortalternative als besonders relevant in Bezug auf den Kontext wahrgenommen wird, sondern auch darauf, welchen Zweck das Modell für eine Schüler_in besitzt. Werner, Schwanewedel und Mayer (2014) zeigen, dass für ihre Untersuchung unterschiedlicher Kontexte bei der Bewertungskompetenz alle Kontext-Personen-Valenzen (z. B. Bekanntheit, Alltags- und Gesellschaftsrelevanz, Interessantheit) positiv mit der Personenfähigkeit korrelierten. In Anlehnung daran kann argumentiert werden, dass der fokussierte Aspekt des biologischen Kontexts nicht in sich selbst leicht oder schwer für Schüler_innen ist, sondern bekannt, relevant oder interessant sein kann und daher von Schüler_innen unterschiedlich bewertet wird. Hieraus lässt sich für die Zusammenstellung der Forced Choice Aufgaben ableiten, dass möglichst Antwortalternativen zusammen präsentiert werden sollten, die sich auf den gleichen Aspekt des Kontexts beziehen, da sonst eine Entscheidung rein nach inhaltlichem Interesse nicht ausgeschlossen werden kann und eine valide Interpretation der Diagnoseergebnisse unmöglich wird (AERA et al., 2014).

7 Fazit und Ausblick

Bei der Entwicklung von Diagnoseaufgaben für Schüler_innen fordern aktuelle Standards der Testentwicklung, die Schüler_innen selbst als Zielstichprobe mehr einzubeziehen (AERA et al., 2014; NRC 2011; Leighton, 2004). Dabei gilt es für eine valide Interpretation der Ergebnisse abzusichern, dass die Schüler_innen das mit dem Test zu messende Wissen und Können auch tatsächlich bei der Bearbeitung der Aufgaben nutzen (AERA et al., 2014). Das in diesem Artikel vorgestellte Vorgehen zur Einbeziehung von Schüler_innen in den Prozess der Aufgabenentwicklung am Beispiel von Diagnoseaufgaben zum Zweck von Modellen machte es möglich, Schwierigkeiten in zuvor konstruierten Antwortalternativen offen zu legen. Die Verständnisprobleme auf Seiten der Schüler_innen würden eine valide zu interpretierende Diagnose unterbinden. Auf der Grundlage der Schülerbewertungen in den Ratingaufgaben und den Schüleraussagen in den Interviews konnten begründet Antwortalternativen selektiert und in Forced Choice Aufgaben zusammengestellt werden.
In einem nächsten Schritt gilt es, die zusammengestellten Forced Choice Aufgaben bei Schüler_innen einzusetzen und empirisch zu überprüfen, ob diese Aufgaben die Hypothese einer validen Diagnose weiter unterstützen. Hierfür werden die Eigenschaften der Aufgaben in einer Studie mittels einer Kombination aus Eyetracking und retrospektivem lautem Denken untersucht (Antwortprozesse; AERA et al., 2014). Eine Dimensionalitätsanalyse wird Aufschlüsse geben, inwieweit die in den Daten ersichtlichen Zusammenhänge zwischen den Aufgaben die theoretisch erwarteten Zusammenhänge bestätigen (Interne Struktur; AERA et al., 2014). Der Vergleich der Diagnoseergebnisse mit externen Variablen und die Überprüfung der Sensibilität der Diagnose bezogen auf Fördermaßnahmen wird die Validität

der Ergebnisinterpretation weiter hinterfragen (Beziehungen zu anderen Variablen und Außenkriterien; AERA et al., 2014).

Literatur

Adams, W. K. & Wieman, C. E. (2011). Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. *International Journal of Science Education, 33* (9), 1289-1312.

AERA, APA & NCME [American Educational Research Association, American Psychological Association & National Council on Measurement in Education] (Hrsg.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cohors-Fresenborg, E., Sjuts, J. & Sommer, N. (2004). Komplexität von Denkvorgängen und Formalisierung von Wissen. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000* (S. 109-144). Wiesbaden: VS Verlag für Sozialwissenschaften.

Crawford, B. A. & Cullin, M. J. (2005). Dynamic assessments of preservice teachers' knowledge of models and modelling. In K. Boersma, M. Goedhart, O. de Jong & H. Eijkelhoff (Hrsg.), *Research and the quality of science education* (S. 309-323). Dordrecht: Springer.

Creswell, J. & Plano Clark, V. (2011). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage.

Ericsson, K. A. & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5* (3), 178-186.

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E. & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft, 16* (1), 5-22.

Gonzalez, E. & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI monograph series: Issues and methodologies in large-scale assessments, 3*, 125-156.

Gorin, J. S. (2007). Test Construction and Diagnostic Testing. In J. Leighton & M. Gierl (Hrsg.), *Cognitive Diagnostic Assessment for Education. Theory and Applications* (S. 173-202). Cambridge: Cambridge University Press.

Grosslight, L., Jay, E., Unger, C. & Smith. (1991). Understanding models and their use in science. Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching, 28* (9), 799-822.

Grünkorn, J. (2014). *Modellkompetenz im Biologieunterricht. Empirische Analyse von Modellkompetenz bei Schülerinnen und Schülern der Sekundarstufe I mit Aufgaben im offenen Antwortformat*. Dissertation.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Harrison, A. G. & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education, 22* (9), 1011-1026.

Hartig, J., Klieme, E. & Leutner, D. (2008). *Assessment of competencies in educational contexts: State of the art and future prospects*. Göttingen: Hogrefe & Huber.

Hodson, D. (2014). Learning Science, Learning about Science, Doing Science. Different goals demand different learning methods. *International Journal of Science Education, 36* (15), 2534-2553.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik*. Weinheim: Beltz.

Justi, R. S. & Gilbert, J. K. (2005). Investigating teachers' ideas about models and modelling: some issues of authenticity. In K. Boersma, M. Goedhart, O. de Jong & H. Eijkelhoff (Hrsg.), *Research and the quality of science education* (S. 325-335). Dordrecht: Springer.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben*. Berlin: Logos.

Kauertz, A., Neumann, K. & Haertig, H. (2012). Competence in Science Education. In B. J. Fraser, K. Tobin & C. J. McRobbie (Hrsg.), *Second International Handbook of Science Education* (S. 711-721). Dordrecht: Springer.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft, 10* (8), 11-29.

KMK [Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland]. (2005). Standards für die Lehrerbildung: Bildungswissenschaften.: Beschluss der Kultusministerkonferenz vom 16.12.2004. *Zeitschrift für Pädagogik, 51* (2), 280-290.

Krell, M., Czeskleba, A. & Krüger, D. (2012). Validierung von Forced Choice-Aufgaben durch Lautes Denken, *Erkenntnisweg Biologiedidaktik, 11*, 53-70.

Krell, M. & Krüger, D. (2013). Wie werden Modelle im Biologieunterricht eingesetzt? Ergebnisse einer Fragebogenstudie, *Erkenntnisweg Biologiedidaktik,12,* 9-26.

Krell, M., Upmeier zu Belzen, A. & Krüger, D. (2012). Students' understanding of the purpose of models in different biological contexts. *International Journal of Biology Education,* 2, 1-34. Verfügbar unter http://www.ijobed.com/2_2/Moritz-2012.pdf

Krell, M., Upmeier zu Belzen, A. & Krüger, D. (2014). Context-specificities in students' understanding of models and modelling: An issue of critical importance for both assessment and teaching. In C. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-Book proceedings of the ESERA 2013 conference. Science education research for evidence-based teaching and coherence in learning. Part 6. Nature of science: History, philosophy and sociology of science.* Nicosia, Cyprus: European Science Education Research Association.

Leighton, J. P. & Gierl, M. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26* (2), 3-16.

Leighton, J. P. (2004). Avoiding Misconception, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice, 23* (4), 6-15.

Mahr, B. (2008). Ein Modell des Modellseins. Ein Beitrag zur Aufklärung des Modellbegriffs. In U. Dirks & E. Knobloch (Hrsg.), *Modelle* (S. 187-218). Frankfurt am Main: Peter Lang.

McCloy, R., Heggestad, E., & Reeve, C. (2005). A silk purse from the sow's ear: retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222-248.

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

NRC [National Research Council]. (2001). *Knowing What Students Know*. Washington, DC: National Academies Press.

Nehm, R. H. & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48* (3), 237-256.

Passmore, C., Gouvea, J. & Giere, R. (2014). Models in science and in learning science. In M. Matthews (Hrsg.), *International handbook of research in history, philosophy and science teaching* (S. 1171-1202). Dordrecht: Springer.

Riemeier, T. (2005). Schülervorstellungen von Zellen, Teilung und Wachstum. *Zeitschrift für Didaktik der Naturwissenschaften, 11* (1), 52-72.

Sandmann, A. (2014). Lautes Denken - die Analyse von Denk-, Lern- und Problemlöseprozesse. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 179-188). Heidelberg: Springer.

Treagust, D. D., Chittleborough, G. D. & Mamiala, T. L. (2002). Students' understanding of the role of scientific models in learning science. *Journal of Science Education, 24* (4), 357-368.

Trier, U. & Upmeier zu Belzen, A. (2009). „Die Wissenschaftler nutzen Modelle, um etwas Neues zu entdecken, und in der Schule lernt man einfach nur, dass es so ist." Schülervorstellungen zu Modellen. *Erkenntnisweg Biologiedidaktik, 8*, 23-37.

Upmeier zu Belzen, A. & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41-57.

Van Driel, J. H. & Verloop, N. (2002). Experienced teachers' knowledge of teaching and learning of models and modelling in science education. *International Journal of Science Education, 24* (12), 1255-1272.

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Werner, M., Schwanewedel, J. & Mayer, J. (2014). Does the context make a difference? Students' abilities in decision-making and the influence of contexts. In C. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-Book proceedings of the ESERA 2013 conference. Science Education Research For Evidence-based Teaching and Coherence in Learning. Part 8. Scientific Literacy and socio scientific issues* (S. 81-89). Nicosia, Cyprus: European Science Education Research Association.

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

Kontakt

Sarah Gogolin
Didaktik der Biologie
Schwendenerstraße 1
14195 Berlin
sarah.gogolin@fu-berlin.de