Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung

Entwicklung eines Testinstruments

Sabrina Mathesius, Annette Upmeier zu Belzen & Dirk Krüger

sabrina.mathesius@fu-berlin.de
Freie Universität Berlin, Didaktik der Biologie,
Schwendenerstraße 1, 14195 Berlin

Zusammenfassung

Das Projekt Ko-WADiS untersucht die Kompetenzen von (Lehramts-)Studierenden der Naturwissenschaften im Bereich der Erkenntnisgewinnung für die Dimensionen „Untersuchen“ und „Modellieren“ fachübergreifend für die Biologie, Chemie und Physik in einer Multi-Kohorten-Längsschnitt-Studie über den Studienverlauf hinweg. Zur Erfassung des Konstrukts wurde theoriegeleitet ein Leistungstest mit Multiple-Choice Aufgaben entwickelt. Die systematische Testkonstruktion wird hier exemplarisch für die Biologie dargestellt und umfasst dabei u. a. die Generierung von Antwortalternativen auf Basis von Studierendenaussagen, die in einer Voruntersuchung (N = 259) qualitativ durch 92 Aufgaben im offenen Format erhoben wurden. Die Analyse der Kennwerte basierend auf den Daten der ersten Erhebungen (N = 3010) zeigt die Eignung von 46 Aufgaben mit biologischen Kontexten. Besonders aufschlussreich sind dabei die Distraktorfunktionskurven, welche eine differenzierte Analyse der Lösungswahrscheinlichkeit im Hinblick auf die Personenfähigkeiten erlauben. Die bereits umgesetzten Aspekte zur Prüfung von Validität, Reliabiliät und Objektivität für die Interpretation der Testergebnisse sollen zukünftig u. a. durch die Erfassung der Lösungsprozesse während der Aufgabenbearbeitung mittels Lauten Denkens ergänzt werden.

Abstract

The Ko-WADiS project aims to model and measure competencies in biology, chemistry, and physics pre-service teacher education, referring in particular to scientific inquiry skills which includes the two inquiry methods „conducting investigations“ and „using scientific models“. A theory-based paper-pencil test has been developed for a longitudinal multi-cohort study. The systematic test development will be presented exemplarily for the items

dealing with biological problems. About 46 of these constructed multiple-choice items using the students' responses (N = 259) to the 92 open-ended tasks as answering options seem to be usable because of the item-parameter analysis (N = 3010). Especially the distractor characteristic curves give detailed information about the item difficulty and the person ability. Additionally aspects of validity, reliability and objectivity will be investigated by taking think-aloud protocols.

1 Einleitung

Der Bereich der Erkenntnisgewinnung bildet einen zentralen Aspekt in der naturwissenschaftlichen Grundbildung und nimmt darüber hinaus eine entscheidende Schlüsselrolle beim wissenschaftlichen Arbeiten innerhalb und außerhalb des Klassenraums ein (BYBEE, 2002; POPPER, 2005). Die Entwicklung eines vertieften Verständnisses der fachmethodischen Prozesse, die zur Generierung von neuen Erkenntnissen führen sowie das Reflektieren über die Charakteristika der Naturwissenschaften sind dabei entscheidend. Nationale Bildungsstandards fordern den Kompetenzerwerb der Schülerinnen und Schüler in diesen Aspekten (z. B. KMK, 2005). Hieraus kann als ein Ziel der Lehrerbildung abgeleitet werden, dass angehende Lehrkräfte im Verlauf ihres Studiums diesbezüglich selbst Kompetenzen aufbauen müssen, um den Unterricht später adäquat gestalten zu können (KMK, 2013). Durch die Entwicklung der professionellen Kompetenz (BAUMERT & KUNTER, 2006) können die Lehrkräfte dazu befähigt werden, Schülerinnen und Schüler angemessen zu unterrichten, zu diagnostizieren und zu fördern.

Während für den schulischen Bildungsbereich bereits Projekte zur Modellierung und Erfassung von Kompetenzen Lernender initiiert wurden (z. B. WELLNITZ & MAYER, 2013), fehlen diese bislang größtenteils für den Hochschulbereich (VON AUFSCHNAITER & BLÖMEKE, 2010). Das Projekt Ko-WADiS[2] (HARTMANN *et al.*, im Druck) begegnet dieser Forschungslücke durch die Modellierung und Erfassung von Kompetenzen (Lehramts-)Studierender im Bereich der naturwissenschaftlichen Erkenntnisgewinnung im Längsschnitt mit Hilfe eines *paper-pencil*-Tests. Ziel des Projekts ist zudem die Untersuchung der Struktur der erhobenen Kompetenzen und die Diskussion der Frage, inwiefern sich Kompetenzunterschiede in Bezug auf verschiedene naturwissenschaftliche Studiengänge und -phasen finden lassen. Kompetenz wird hier als kogni-

2 Das Akronym Ko-WADiS steht für: *Ko*mpetenzmodellierung und -erfassung zum *W*issenschaftsverständnis über naturwissenschaftliche *A*rbeits- und *D*enkweisen be*i S*tudierenden (Lehramt) in den drei naturwissenschaftlichen Fächern Biologie, Chemie und Physik. Wir danken dem Bundesministerium für Bildung und Forschung, dass es Ko-WADiS im Rahmen des Forschungsprogramms KoKoHs fördert.

tive Leistungsdisposition aufgefasst (HARTIG & KLIEME, 2006). Neben Lehramtsstudierenden der Fächer Biologie, Chemie und Physik werden vergleichend Studierende der naturwissenschaftlichen Studiengänge ohne diese didaktische Ausrichtung sowie Studierende an österreichischen Universitäten, deren Studium bislang nicht im Zuge der Bologna-Reform entsprechend kompetenzorientiert modularisiert wurde, befragt.[3]

2 Theoretische Anknüpfung

Naturwissenschaftliche Erkenntnisgewinnung kann als komplexer Problemlöseprozess verstanden werden, der bezugnehmend auf das Rahmenkonzept wissenschaftsmethodischer Kompetenzen die Anwendung von methodischem und inhaltlichem Konzeptwissen erfordert (KLAHR, 2000; vgl. MAYER, 2007). Im Projekt wird die Verknüpfung des Standards der Erkenntnisgewinnung *scientific inquiry* und des Kompetenzkonstrukts *scientific reasoning* betrachtet. Als Schnittmenge aller drei Naturwissenschaften und aufgrund ihrer Bedeutung im wissenschaftlichen Kontext wird hierbei auf die Arbeits- und Denkweisen „Untersuchen“ (MAYER, 2007) und „Modellieren“ (UPMEIER ZU BELZEN & KRÜGER, 2010) fokussiert (Ableitung von Indikatoren siehe Tab. 1). Dabei stützen sich die Überlegungen vor allem auf bereits bekannte Befunde für die Kompetenzen von Lehrkräften (z. B. CRAWFORD & CULLIN, 2005; KUNZ, 2010) sowie von Schülerinnen und Schülern (z. B. WELLNITZ & MAYER, 2013).

2.1 Instrumente zur Kompetenzmessung

Im Zuge der Umstrukturierung von schulischen und universitären Bildungsgängen rückt die Kompetenzmessung vermehrt in den Fokus der Forschung. Hierfür wurden u. a. für den Kompetenzbereich Erkenntnisgewinnung bereits einzelne Struktur- und Niveaumodelle konzipiert und evaluiert (z. B. WELLNITZ & MAYER, 2013). Dabei wurden teilweise Aufgaben im offenen Antwortformat eingesetzt, welchen eine größere Validität zugesprochen wird, jedoch bei gleichzeitig erhöhter Auswertungsdauer im Vergleich zu Aufgaben im geschlossenen *Multiple-Choice* (*MC*)-Format (HARTIG & JUDE, 2007). Weitere eingesetzte Testformate bilden *hands-on* Aufgaben (z. B. ORSENNE & UPMEIER ZU BELZEN, 2012) und computerbasierte Lernumgebungen (z. B. SCHERER &

3 Projektbeteiligte sind die Fachdidaktiken Biologie und Physik der Freien Universität Berlin sowie die Fachdidaktik für Lehr-/Lernforschung Biologie und die Fachdidaktik Chemie der Humboldt-Universität zu Berlin ebenso wie Kooperationspartner an den beteiligten Universitäten (Universität Duisburg-Essen, Universität Innsbruck, Universität Salzburg und Universität Wien) der Vergleichsstichproben.

TIEMANN, 2012). Beide Methoden fokussieren vermehrt die Durchführung des Modellierens bzw. Untersuchens. Das Testformat kann die kognitiven Anforderungen an die Beantwortung beeinflussen und steht im engen Zusammenhang mit Testgütekriterien (MARTINEZ, 1999).

In der vorliegenden Multi-Kohorten-Längsschnitt-Studie wird an jeweils vier Messzeitpunkten eine Vollerhebung der Kompetenzen der Lehramtsstudierenden mindestens eines naturwissenschaftlichen Faches an den beteiligten Berliner Universitäten angestrebt ($n_{\text{Kohorte}} \approx 400$). Unter Berücksichtigung der Testökonomie wird der Einsatz von Aufgaben im geschlossenen *MC*-Format bevorzugt. Zur Untersuchung der Kompetenzen von Studierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung mittels *MC*-Aufgaben empfiehlt es sich, zunächst qualitativ Perspektiven der Probanden zu erfassen, auf deren Basis die Aufgaben schrittweise weiterentwickelt werden (SADLER, 1998). Durch diese Art des Konstruktionsprozesses wird der unsicheren Inhaltsvalidität von geschlossenen Aufgaben begegnet (HARTIG & JUDE, 2007). Zudem kann die Ratewahrscheinlichkeit durch authentische, plausible Distraktoren und Attraktoren minimiert werden, was einen positiven Einfluss auf die Reliabilität und kriteriale Validität der Aufgaben haben kann.

2.2 Das „Untersuchen“ als Teil der Erkenntnisgewinnung

Naturwissenschaftliche Untersuchungen können nach dem hypothetisch-deduktiven Verfahren der Erkenntnisgewinnung theoretisch beschrieben werden (POPPER, 2005). Als Untersuchungen werden im hier vorgestellten Projekt die Arbeitsweisen „Beobachten“ (korrelative Zusammenhänge) und „Experimentieren“ (kausale Zusammenhänge) verstanden (WELLNITZ & MAYER, 2013). Es lassen sich vier Prozessvariablen des wissenschaftlichen Denkens in jeweils fünf Niveauausprägungen unterscheiden: naturwissenschaftliche Fragestellungen generieren, Hypothesen generieren, Untersuchungen planen sowie Daten analysieren und Schlussfolgerungen ziehen. Wichtige Aspekte sind dabei der Umgang mit unabhängiger und abhängiger Variable sowie Kontroll- bzw. Störvariablen; des Weiteren spielen Messzeiten und Messwiederholungen eine Rolle (WELLNITZ & MAYER, 2013). Studien belegen erhebliche Defizite bei Lehrkräften in allen Bereichen der Prozessvariablen und weisen dabei auf vorherrschende Perspektiven hin, die im Unterricht berücksichtigt und gefördert werden sollten (z. B. KUNZ, 2010).

2.3 Das „Modellieren“ als Teil der Erkenntnisgewinnung

Der Einsatz von Modellen ist vielfältig, so dass Modelle zum einem als Medium genutzt werden können (Herstellungsperspektive), aber das Modellieren

zum anderen auch als Methode und somit als naturwissenschaftliche Arbeitsweise verstanden werden kann (Anwendungsperspektive; vgl. MAHR, 2008). Beide Perspektiven eröffnen dabei unterschiedliche Anforderungen an die Lernenden. Studien zeigen, dass Lehrkräfte Modelle selten aus der methodischen Anwendungsperspektive heraus verstehen und somit Modelle zur Generierung neuer Erkenntnisse auch nur wenig im naturwissenschaftlichen Unterricht einsetzen (z. B. CRAWFORD & CULLIN, 2005). Das Modellieren als Modellbildungsprozess kann dabei untergliedert werden in den Zweck, das Testen und das Ändern von Modellen (UPMEIER ZU BELZEN & KRÜGER, 2010).

3 Fragestellungen

Für den Entwicklungsprozess eines Messinstruments zur Erfassung von Kompetenzen im Bereich der Erkenntnisgewinnung für die Dimensionen „Untersuchen“ und „Modellieren“ bei Studierenden unterschiedlicher naturwissenschaftlicher Studiengänge und -fächer ergeben sich folgende Forschungsfragen:

F1: Inwiefern gelingt es, durch Aufgaben im offenen Antwortformat generierte Studierendenaussagen zur Entwicklung eines Testinstruments im geschlossenen Format zu nutzen, dessen Messergebnisse als valide interpretiert werden können?

F2: Inwieweit erfüllen *MC*-Aufgaben, die mit Hilfe von Probandenaussagen entwickelt wurden, Gütekriterien für Testaufgaben?

Die Prüfung der Validität ist die Beurteilung der Testergebnisse im Hinblick auf das zu messende Konstrukt; sie setzt Objektivität und Reliabilität voraus (SCHMIEMANN & LÜCKEN, 2014). Dabei ist nach aktuellem Forschungsdiskurs Validität nicht als ein reines Testmerkmal zu verstehen, vielmehr wird die Validität von differenzierten Interpretationen der Messergebnisse beispielsweise für Inhalt, Kriterium und Konstrukt in den Fokus genommen (MESSICK, 1995). Das Einbeziehen von Studierendenperspektiven in den Entwicklungsprozess eines geschlossenen Leistungstests kann zur Validität der Interpretation der Testergebnisse beitragen (SADLER, 1998). Der Erfolg dieses Konstruktionsprozesses lässt sich durch die generierten Itemkennwerte prüfen. Die Beachtung verschiedener Qualitätsschritte im Entwicklungsprozess lässt zudem die psychometrische Eignung der konstruierten Aufgaben erwarten (TERZER, HARTIG & UPMEIER ZU BELZEN, 2013).

4 Methodisches Vorgehen

Der Konstruktionsprozess für das zu entwickelnde Testinstrument wird hier für die Domäne Biologie vorgestellt (Abb. 1). Die im Projekt angestrebte Betrachtung der Entwicklungsprozesse bei Studierenden bezieht sich dabei auf die Populationsebene, d. h., es sollen zunächst keine individuellen Messergebnisse betrachtet werden. Mit Blick auf das Multi-Kohorten-Längsschnitt-Design der Studie wurde zugunsten der Testökonomie und der erhöhten Auswertungsobjektivität ein *paper-pencil*-Test im geschlossenen Aufgabenformat entwickelt (MOOSBRUGGER & KELAVA, 2012). Dies erlaubt gegenüber der Integration von stammäquivalenten Aufgaben im offenen Format zudem die Beantwortung einer höheren Itemanzahl pro Proband innerhalb des Bearbeitungszeitraums.

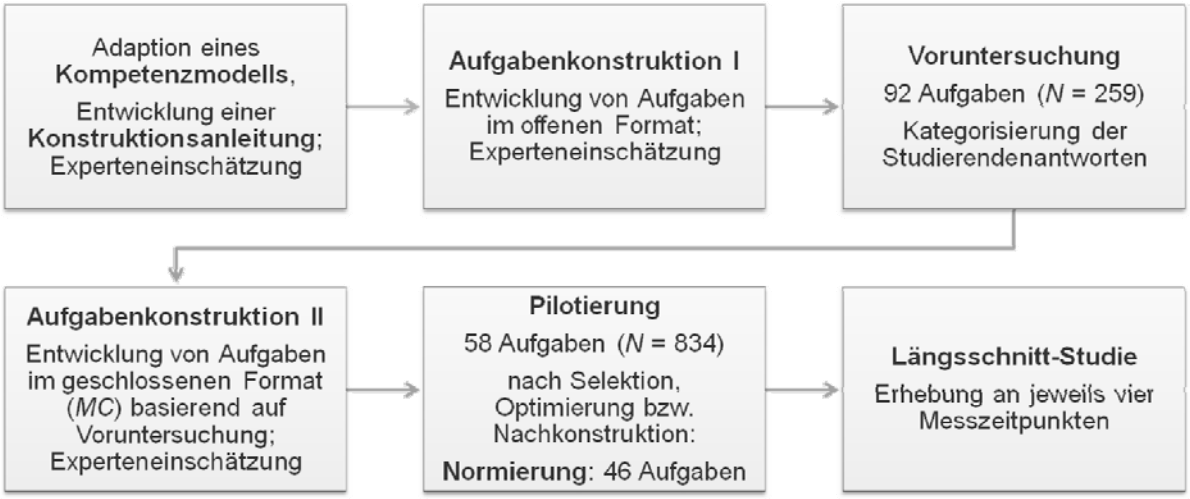

Abbildung 1: Prozess der Testentwicklung (verändert nach: TERZER *et al.*, 2013).

4.1 Adaption eines Kompetenzmodells und Entwicklung einer Konstruktionsanleitung

Basierend auf theoretisch hergeleiteten Kompetenzmodellen zum „Untersuchen“ und „Modellieren“ wurden Indikatoren für das zu messende Konstrukt erarbeitet (Tab 1). Dabei liegt der Schwerpunkt auf einer elaborierten Sichtweise: dies beinhaltet eine gleichzeitige Integration verschiedener Anforderungen an naturwissenschaftliche Untersuchungen sowie den Umgang mit naturwissenschaftlichen Modellen unter methodischer und nicht medialer Perspektive.

Zur Standardisierung der Schritte Aufgabenkonstruktion I und II (Abb. 1) wurde eine Anleitung mit Operationalisierungen für die einzelnen Teildimensionen konzipiert und deren Passung zum Konstrukt durch Experten aus den Naturwissenschaftsdidaktiken und der Psychologie beurteilt (vgl. MOOSBRUGGER & KELAVA, 2012). Die Anleitung enthält Hinweise sowohl für die Aufga-

ben im offenen als auch für die Aufgaben im geschlossenen Format; u. a. Beispiele zum generellen Aufbau (Stamm, standardisierter Impuls) sowie für die Aufgaben im offenen Format einen theoretisch hergeleiteten Erwartungshorizont. Der Aufgabenstamm bildet stets ein naturwissenschaftliches Problem ab, welchem die Probanden begegnen, und beinhaltet ggf. Abbildungen, Graphen oder Tabellen. Das benötigte inhaltliche Fachwissen ist dabei gegeben, so dass allein die fachmethodischen Kompetenzen zur Lösung herangezogen werden müssen (Beispiel für beide Konstruktionsschritte siehe Tab. 3).

Tabelle 1: Anforderungen in den *MC*-Aufgaben bezogen auf die Dimensionen „Untersuchen“ (vgl. Mayer, 2007) sowie „Modellieren“ (vgl. UPMEIER ZU BELZEN & KRÜGER, 2010) und die entsprechenden Teildimensionen als Indikatoren des Konstrukts.

Dimension	Teildimension	Anforderungen in den Aufgaben
		In dieser Teilkompetenz beurteilen Studierende, inwieweit…
Untersuchen	Fragestellungen formulieren	… naturwissenschaftliche Fragestellungen einen Bezug zu Phänomen haben, empirisch überprüfbar, intersubjektiv nachvollziehbar, eindeutig, grundsätzlich beantwortbar und intern sowie extern konsistent sind.
	Hypothesen generieren	… Hypothesen empirisch überprüfbar, intersubjektiv nachvollziehbar, eindeutig, logisch widerspruchsfrei und in Vereinbarkeit mit einer zugrundeliegenden Theorie sind.
	Planen von Untersuchungen	… kausale Zusammenhänge zwischen unabhängiger und abhängiger Variable aufgrund einer vorhergehenden Hypothese untersucht werden, wobei die unabhängige Variable beim Experimentieren gezielt manipuliert wird und Kontrollversuche sowie Störvariablen in den Blick genommen werden. … korrelative Zusammenhänge zwischen unabhängiger und abhängiger Variable aufgrund einer vorhergehenden Hypothese in einer Beobachtung untersucht werden.
	Auswertung von Untersuchungen	… durch die Auswertung von Rohdaten eine Interpretation und Reflexion bezüglich der aufgestellten Forschungsfrage und Hypothese möglich wird.
Modellieren	Zweck von Modellen	… Modelle zur Generierung von Hypothesen genutzt werden.
	Testen von Modellen	… sich mit Hilfe von Modellen aufgestellte Hypothesen prüfen lassen.
	Ändern von Modellen	… Modelle aufgrund von Falsifikation der durch sie aufgestellten Hypothesen geändert werden.

4.2 Aufgabenkonstruktion I und Voruntersuchung

Im Schritt Aufgabenkonstruktion I (Abb. 1) wurden auf der Basis einer Konstruktionsanleitung Aufgaben im offenen Format entwickelt ($N = 92$), die für alle Teildimensionen ein biologisches Problem in den Fokus stellen. Die Auswahl der Inhalte für die Aufgabenkontexte basiert auf einer Analyse des Curriculums für den Studiengang Biologie sowie den Studiengang Biologie mit Lehramtsoption und bezieht sich auf die Modulbeschreibungen der Studienord-

nungen der kooperierenden Universitäten des Wintersemesters 2011/2012. Bei der Konzeption der Aufgabenkontexte wurde somit auf eine inhaltlich breite Streuung geachtet, welche für die fachmethodischen Anforderungen relevante Situationen darstellen. Die fachliche Korrektheit und Passung zur Konstruktionsanleitung wurde durch Expertenurteile überprüft.

Die im offenen Antwortformat entwickelten Aufgaben aus Konstruktionsprozess I wurden in einer Voruntersuchung bei 259 Studierenden mit mindestens einem naturwissenschaftlichen Fach an einer der Berliner Universitäten eingesetzt; die Stichprobe stellt somit eine Zufallsstichprobe auf Basis der Gesamtstichprobe dar. Zu Beginn einer jeden Testung wurde zur Wahrung der Durchführungsobjektivität eine standardisierte Einführung in das Forschungsprojekt mit einer Erläuterung des Vorgehens bei der Befragung gegeben. Die Teilnahme an der Studie war freiwillig und die Daten wurden anonym erhoben. Allgemeine soziodemografische Daten[4] wurden ergänzend erfasst.

Die jeweils pro Proband bearbeiteten 14 bis 15 Aufgaben wurden auf 20 Testhefte im Multi-Matrix-Design verteilt. Ergänzend wurden Fragen zur Beurteilung der Aufgaben durch die Probanden gestellt. Die Fragen befassen sich mit der Sicherheit und Anstrengung der Probanden bezüglich der Beantwortung und erfassen, wie hilfreich für die Lösung das vorhandene Fachwissen, Methodenkenntnisse bzw. das logische Schlussfolgern waren. Die Einschätzung erfolgte mittels einer sechsstufigen Likert-Skala. Die Studierenden wurden zudem dazu angehalten, für sie Unverständliches sowie generelle Anmerkungen direkt im Fragebogen zu notieren. Bei der Kategorisierung der schriftlichen Antworten wurde die Software MAXQDA 10 genutzt.

4.3 Aufgabenkonstruktion II und Pilotierung

Von den erhobenen Studierendenaussagen ausgehend wurden Antwortalternativen für *MC*-Aufgaben entwickelt (Aufgabenkonstruktion II; Abb. 1). Die Formulierung der Antwortalternativen entspricht dabei generellen Richtlinien zur Konstruktion von Testaufgaben (JONKISZ, MOOSBRUGGER & BRANDT, 2012). Es wurde darauf geachtet, dass die Formulierungen nicht zu kompliziert, möglichst gleich lang und syntaktisch gleich aufgebaut sind. Die sprachliche und inhaltliche Verständlichkeit und Komplexität wurde mit Experten diskutiert, Fremdwörter wurden vermieden; zudem wurde die generelle Anbindung an die Sprache und Lebenswelt der Studierenden und die gleichmäßige Ver-

4 Soziodemografische Daten: Geschlecht, Alter, Universitätszugehörigkeit, studierte Fächerkombination, Studiensemesterzahl, Art des Studiengangs und Abschlussziel (Lehramt ja/ nein).

wendung von Schlüsselbegriffen in allen Antwortalternativen beachtet (NEUHAUS & BRAUN, 2007). Bei der Reihenfolge der Antwortalternativen spielen inhaltliche Aspekte keine Rolle, so dass diese zufällig angeordnet wurden; zugleich wurden wechselseitige Beziehungen zwischen den Aussagen ausgeschlossen (JONSKISZ *et al.*, 2012). Inwiefern die konstruierten Antwortalternativen das adaptierte Kompetenzmodell operationalisieren, wurde durch Expertenurteile bezogen auf einen möglichen Repräsentationsschluss abgesichert.

Die konstruierten *MC*-Aufgaben für die drei Fächer wurden zu Blöcken zusammengefasst (je Block 3 Aufgaben eines Fachs/Dimension), welche mittels Multi-Matrix-Design zu 49 Testheften mit jeweils 18 bis 24 Aufgaben sortiert wurden. Für die Aufgaben mit biologischen Kontexten bestehen keine logischen Abhängigkeiten, so dass eine freie Anordnung möglich ist. Die Blöcke wurden an unterschiedliche Positionen im Testheft rotiert, um Positionseffekte zu vermeiden (je Testheft mindestens ein Block je Fach und je Dimension).

In der ersten Untersuchungsperiode (Sommersemester 2013, Wintersemester 2013/14) wurden die selektierten Aufgaben bei $N = 3010$ Probanden (54.4% weiblich; Alter: $MW = 22.59$, $SD = 4.41$; 1745 Biologiestudierende) an 11 beteiligten Universitäten[5] eingesetzt. Die Stichprobe setzte sich zu 54.7 % aus Lehramtsstudierenden und zu 45.3 % aus Studierenden eines Mono-Studiengangs zusammen; 2144 Bachelor-Studierende wurden befragt.

Nach dem Einsatz der entwickelten Testaufgaben bei Studierenden wurden auf der Basis von Kennwerten Aufgaben zur längsschnittlichen Erfassung der Kompetenzen ausgewählt (vgl. TERZER *et al.*, 2013). Zur Einschätzung der Eignung der *MC*-Aufgaben wurden unterschiedliche Merkmale herangezogen: Itemschwierigkeit, Itemtrennschärfe, *Item fit* (*wMNSQ*, *T*-Wert) und charakteristische Distraktorfunktionskurven. Die Itemschwierigkeit wird durch IRT-Analysen mittels Rasch-Skalierung (Itemparameter) und auf Basis der klassischen Lösungswahrscheinlichkeit betrachtet. Um eine ausreichende diagnostische Funktion zu gewährleisten, sollten Aufgaben mit stark verminderter bzw. erhöhter Lösungswahrscheinlichkeit vom Test ausgeschlossen werden. Mittels der Itemtrennschärfe gelingt es einzuschätzen, inwiefern die Items dazu geeignet sind, zwischen kompetenten und weniger kompetenten Probanden zu unterscheiden und somit das zu messende Konstrukt angemessen abzubilden (KELAVA & MOOSBRUGGER, 2012). Gleichzeitig bietet sie Hinweise für die Reliabilität des Tests. Hierfür werden alle Aufgaben mit ConQuest 3.0

5 Beteiligte Universitäten: Freie Universität Berlin, Humboldt-Universität zu Berlin, Universität Duisburg-Essen, Technische Universität Berlin, Universität Potsdam, RWTH Aachen, Universität Bremen, Universität zu Köln, Universität Innsbruck, Universität Wien und Universität Salzburg.

(ADAMS, WU & WILSON, 2012) eindimensional skaliert, wodurch die Trennschärfe in Relation zu allen Testaufgaben gesetzt wird. Die Beurteilung der *Item fit*-Werte sollte bei einem Erwartungswert von 1.00 bei $0.70 \leq wMNSQ \leq 1.30$ (BOND & FOX, 2007) liegen. Den Zusammenhang von Lösungswahrscheinlichkeit und Personenfähigkeit zeigen die charakteristischen Item- bzw. Distraktorfunktionskurven für erwartete und beobachtete Messwerte an, wobei eine Angleichung von theoretisch angenommener und empirischer Kurve für eine gute Itemqualität angestrebt wird.

5 Ergebnisse

Es wurden 92 Aufgaben im offenen Antwortformat konstruiert (Tab. 2). Inhaltlich fächern die gewählten Kontexte für die Aufgaben breit auf und streuen in verschiedenen Schwierigkeiten. Für die Teildimension „Fragestellungen formulieren“ lässt sich beispielsweise erkennen, dass die Schwierigkeit der Aufgabe steigt, je allgemeiner die zu generierende Fragestellung ist (vgl. GRUBE, 2010).

Tabelle 2: Anzahl und Verteilung der Aufgaben für die Konstruktionsschritte I (offenes Antwortformat) und II (geschlossenes Antwortformat, *MC*).

Dimension	**Teildimension**			
Untersuchen	Fragestellungen formulieren	Hypothesen generieren	Planung von Untersuchungen	Auswertung von Untersuchungen
Konstruktion I | II	9 | 9	9 | 8	21 | 11	8 | 9
Modellieren	Zweck von Modellen	Testen von Modellen	Ändern von Modellen	
Konstruktion I | II	15 | 7	14 | 8	16 | 6	

Die Kategorisierung der 10-20 Aussagen pro Aufgabe ließ eine Zuordnung in fachlich richtig und falsch zu und erlaubte aufgrund der Kategorisierung in unterschiedliche Perspektiven die Generierung von Antwortalternativen für den Konstruktionsprozess II (Tab. 3). Die Probandenaussagen zeigen dabei eine unterschiedliche Qualität und ermöglichen teilweise eine Kategorisierung anhand bekannter Schwierigkeiten beim „Untersuchen“ und „Modellieren“; z. B. lässt sich ein unsystematischer Umgang mit Variablen (Zuordnung von abhängiger/ unabhängiger Variable innerhalb der Teildimension „Planung von Untersuchungen“) erkennen (vgl. HAMMANN, 2004); teilweise wurden neue Kategorien induktiv gebildet.

Tabelle 3: Verworfene Aufgabe „Vegetation auf Wanderwegen“ (Teildimension „Fragestellungen formulieren“) mit ausgewählten Studierendenaussagen, induktiv abgeleiteten Kategorien ([…]) und daraus konstruierten Antwortalternativen der *MC*-Aufgabe; n = Probandenanzahl, p_i = Lösungswahrscheinlichkeit, r_i = Trennschärfe.

Aufgabe	Offenes Antwortformat	Geschlossenes Antwortformat
Stamm	In Wäldern und auf Wiesen gibt es oft kleine „Trampelpfade“, auf denen trotz üppiger Vegetation der restlichen Umgebung wenig Pflanzenwachstum zu beobachten ist.	
Impuls	*Formulieren Sie eine naturwissenschaftliche Fragestellung, die sich aus diesem Phänomen ableiten lässt.*	*Welche naturwissenschaftliche Fragestellung lässt sich aus diesem Phänomen ableiten? Kreuzen Sie an!*
Antworten	Welche Faktoren beeinflussen das Wachstum von Pflanzen? [*Attraktor*]	Welche Faktoren beeinflussen das Wachstum von Pflanzen auf Wanderwegen? [n_A = 13]
	Welcher Zusammenhang besteht zwischen dem Niedertrampeln und dem Absterben der Vegetation? [*bereits bekannter Faktor*]	Welche Korrelation besteht zwischen dem Benutzen von Wanderwegen und dem Absterben von Pflanzen? [n_B = 26]
	Welche Auswirkungen hat das menschliche Eingreifen in die Natur durch Trampelpfade auf die Vegetation eines Biotops? [*irrelevanter Faktor*]	Welche Auswirkungen auf die Vegetation hat das menschliche Eingreifen? [n_C = 6]
	Beeinflusst Wildwechsel maßgeblich die Vegetation eines Biotops? [*zu weit greifend*]	Welchen Einfluss üben Lebewesen auf die Vegetation von Wäldern und auf Wiesen aus? [n_D = 19]
Parameter	p_i = 0.2 | *wMNSQ* = 1.00 | T = 0.01 | r_i = 0.16 | N = 64	

Die konstruierten 58 *MC*-Aufgaben mit biologischen Kontexten wurden im Rahmen mehrerer Prä-Pilotierungen auf Verständlichkeit und psychometrische Qualität geprüft; die Anzahl reduzierte sich daraufhin auf 46 Aufgaben für die Biologie (141 für alle Fächer), welche im Mittel von 359 Probanden (*SD*=138) bearbeitet wurden. Diese verbleibenden Aufgaben zeigen zufriedenstellende *Item fit*-Werte in folgenden Bereichen: $0.95 \leq wMNSQ \leq 1.06$; $-1.9 \leq T \leq 2$; die korrigierte Trennschärfe liegt zwischen $0.11 \leq r_i \leq 0.46$. Die Analyse der Parameter (mittlere Schwierigkeit = 0.28 logits) zeigt zudem, dass die Personenfähigkeiten (Mittelwert zur Parameterschätzung auf null fixiert) gut durch Aufgaben mit unterschiedlichen Schwierigkeiten abgedeckt sind; die Lösungswahrscheinlichkeiten liegen zwischen 0.19 und 0.8. Die EAP/PV-Reliabilität beträgt 0.47. Die Distraktorenanalyse für die ausgewählten Aufgaben zeigt eine Anwahl-wahrscheinlichkeit der Distraktoren zwischen 2% und 55%. Die grafische Darstellung der Wahrscheinlichkeit für die Wahl der Antwortoptionen in Abhängigkeit von der Personenfähigkeit zeigt innerhalb der aktuellen Analysen der 46 Aufgaben Distraktorfunktionskurven mit guter und

weniger guter Passung zu den Erwartungswerten (Abb. 2 zeigt eine gute Passung).

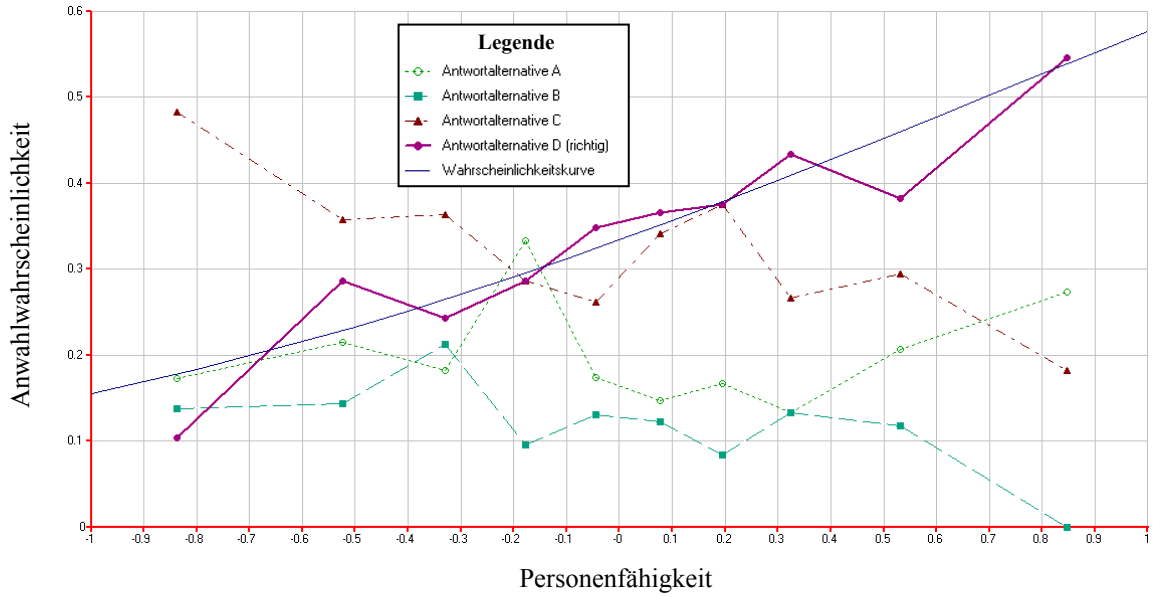

Abbildung 2: Distraktorfunktionskurve für die Aufgabe „Fledermäuse und Pflanzen“ zeigt, dass mit zunehmender Personenfähigkeit die Attraktivität der Distraktoren abnimmt, besonders fähige Personen wählen den Attraktor (Teildimension „Hypothesen generieren“; p_i = 0.35, *wMNSQ* = 0.99, T = -0.1, r_i = 0.29, Delta = 0.69, N = 307).

6 Diskussion

Innerhalb der Testentwicklung wurden verschiedene Maßnahmen umgesetzt, um einzuschätzen, inwiefern die erhaltenen Testergebnisse als valide in Bezug auf die Inhalts- und Kriteriumsvalidität betrachtet werden können (F1).

Aufgrund der Orientierung an Studienordnungen war es möglich, sowohl inhaltlich als auch bezogen auf die Schwierigkeit eine breite Auffächerung der Kontexte zu realisieren und eine möglichst große Merkmalsausprägung im Bereich der naturwissenschaftlichen Erkenntnisgewinnung abzubilden (JONKISZ *et al.*, 2012). Allerdings sind nicht alle Zellen der untersuchten Teildimensionen mit der gleichen Aufgabenanzahl besetzt (vgl. Tab. 2), so dass keine Schätzung auf Teildimensionsebene erfolgen sollte.

Durch das generelle Verständnis des Stamms und des standardisierten Impulses der Aufgaben im offenen Format konnten die Studierenden zu differenzierten Antworten angeregt werden. Die Probandenaussagen zeigten eine Vielzahl unterschiedlicher Perspektiven, welche entsprechenden Kategorien zugeordnet wurden. Es konnte dabei an bereits vorhandene Beschreibungen zum Fehlen einer elaborierten Sicht auf das Modellieren (CRAWFORD & CULLIN, 2005) und bezogen beispielsweise auf die Kontrolle von Variablen in

einem zu entwickelnden Setting (HAMMANN, 2004) angeknüpft werden und es wurden auch induktiv neue Kategorien gebildet (z. B. Tab. 3). Auf Basis einzelner Kategorien wurden Aufgaben im geschlossenen Format mit einem Attraktor sowie drei Distraktoren entwickelt. Stämme, aufgrund derer keine reichhaltigen Antworten generiert werden konnten, wurden verworfen bzw. aufgrund von Anmerkungen der Studierenden überarbeitet und erneut in einer Voruntersuchung geprüft (Abb. 1). Auf diese Weise ist es früh im Entwicklungsprozess möglich, geeignete Aufgabenstämme zu identifizieren.

Die Analyse der Itemkennwerte für die konstruierten *MC*-Aufgaben erlaubt weitere Rückschlüsse zur Einschätzung des Testinstruments (F2). Insgesamt lässt sich erkennen, dass die konstruierten Antwortalternativen für alle Aufgaben durch Probanden angewählt wurden, so dass diese eine gewisse Plausibilität und Attraktivität haben (SADLER, 1998; vgl. Distraktoranwahlwahrscheinlichkeit). Es zeigt sich zudem eine gute Abdeckung der Personenfähigkeiten durch Aufgaben mit unterschiedlichen Schwierigkeiten. Die mittlere Aufgabenschwierigkeit liegt dabei leicht über der mittleren Personenfähigkeit, so dass angenommen werden kann, dass bei der längsschnittlichen Untersuchung mit dem konstruierten Instrument keine Deckeneffekte auftreten. Da Studierende aller Semester, welche mindestens ein naturwissenschaftliches Fach belegen, mit den *MC*-Aufgaben querschnittlich befragt wurden, konnte daher bereits während des Entwicklungsprozesses die Einsetzbarkeit des Testinstruments über die Studienart und Semesteranzahl betrachtet werden.

Bei der Selektion von Aufgaben auf Basis der Distraktorfunktionskurven kann eine adäquate Abbildung des Konstrukts erfolgen, wenn besonders fähige Personen den Attraktor auswählen, wohingegen Distraktoren vermehrt gewählt werden, wenn eine geringere Personenfähigkeit vorliegt (WALPUSKI & ROPOHL, 2014; vgl. Abb. 2). So konnte auf Basis der Pilotierung bereits eine dezidierte Betrachtung der Distraktorenauswahl erfolgen und es ließen sich dadurch Hinweise für eine selektive Optimierung sammeln, indem Aufgaben mit weniger erwartungskonformen Kurvenverläufen verworfen wurden.

Ein weiteres Kriterium zur Einschätzung der Validität der Testergebnisse und somit auch zur konformen Abbildung des Konstrukts kann die Trennschärfe sein, welche anzeigt, inwiefern eine Probandenantwort sich zum Antwortverhalten im gesamten Test verhält. Aufgaben mit zu geringer Trennschärfe wurden vom Test entfernt (vgl. Trennschärfe für die Aufgabe in Tab. 3).

Die Analysen auf Ebene der einzelnen Aufgaben werden durch die Einschätzung auf Testebene ergänzt, so dass eine Modellselektion erfolgen kann. Indizien für ein eindimensionales Konstrukt stützen beispielsweise die Annah-

me, dass es sich bei den fokussierten Kompetenzen der Erkenntnisgewinnung unter Umständen um für ähnliche Situationen generalisierbare Ausprägungen handelt (HARTIG & KLIEME 2006; MAYER 2007; vgl. HARTMANN *et al.*, im Druck). Dies wird in ein- und mehrdimensionalen Analysen weiter untersucht. Hierbei könnten aufgrund engerer Parametergrenzen teilweise weitere Aufgaben ausgeschlossen werden; die aktuell 46 selektierten Aufgaben weisen *Item fit*-Werte in einem akzeptablen Bereich auf (BOND & FOX, 2007).

7 Fazit und Ausblick

Es konnte gezeigt werden, dass der Entwicklungsprozess eines Messinstruments mit *MC*-Aufgaben auf Basis von erhobenen Probandenaussagen Vorteile von quantitativen und qualitativen Aufgabenformaten verbindet (SADLER, 1998) und somit einen Beitrag zur Validität der Interpretation der Testergebnisse leistet (MESSICK, 1995; SCHMIEMANN & LÜCKEN, 2014). Der Validitätsaspekt kann weiter untersucht werden, indem der Frage nachgegangen wird, welche kognitiven Prozesse durch die Beantwortung der Aufgaben initiiert werden und inwieweit diese das abzubildende Konstrukt repräsentieren. In einer anknüpfenden Studie werden Studierende mit verschiedenen Fächerkombinationen und auf Basis des Messinstruments geschätzten unterschiedlichen Kompetenzausprägungen mittels Lauten Denkens (ERICSSON & SIMON, 1980) dazu aufgefordert, ihre Lösungsprozesse bei der Bearbeitung zu verbalisieren (vgl. TERZER, PATZKE & UPMEIER ZU BELZEN, 2012).

Der weitere Einsatz der selektierten *MC*-Aufgaben in der Multi-Kohorten-Längsschnitt-Studie wird je Kohorte an vier Messzeitpunkten stattfinden. Neben den bereits beschriebenen Aspekten zur Validitätsprüfung für die Messergebnisse des Testinstruments (MESSICK, 1995) aufgrund der curricularen Verankerung, der Konstruktionsanleitung inklusive Operationalisierung und Experteneinschätzung (Inhaltsvalidität) werden weitere Schritte zukünftig strukturelle Aspekte anhand spezifischer Dimensionalitätsprüfungen beinhalten. Zudem wird in einem latenten Regressionsmodell der Einfluss bestimmter Personenmerkmale durch die erhobenen Hintergrundvariablen untersucht. Der externen Validitätsprüfung wird durch eine diskriminante Erfassung der kognitiven Grundfähigkeit und konvergent durch den Einsatz von Aufgaben eines Messinstruments zum Professionswissen von Lehramtsstudierenden begegnet. Des Weiteren sind in Teilprojekten vergleichende Analysen hinsichtlich der Vorstellungen (*beliefs)* im Bereich der Charakteristika der Naturwissenschaften (*Nature of Science)* und der allgemeinen Problemlösefähigkeit geplant. Mit Hilfe des überarbeiteten Messinstruments sollen die Kompetenzen von Studie-

renden quer- und längsschnittlich erfasst werden und es wird dabei angestrebt, einen Beitrag zur Analyse und Optimierung von universitären Bildungsprozessen innerhalb der Lehrerbildung zu initiieren. Hierfür sollten zukünftig ergänzend auch die Lerngelegenheiten für fachmethodische Kompetenzen der Erkenntnisgewinnung während des Studienverlaufs in den Blick genommen werden.

Zitierte Literatur

ADAMS, R. J., WU, M. L., & WILSON, M. R. (2012). *Conquest 3.0 [computer software]*. Camberwell: Australian Council for Educational Research.

BAUMERT, J., & KUNTER, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*, 469–520.

BOND, T. B., & FOX, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences.* Mawah, NJ: Lawrence Erlbaum.

BYBEE, R. W. (2002). Scientific Literacy – Mythos oder Realität. In W. GRÄBER, P. NENTWIG, T. KOBALLA, & R. EVANS (Hrsg.), *Scientific Literacy. Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung* (S. 21-43). Opladen: Leske + Budrich.

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. EIJKELHOF (Hrsg.), *Research and the Quality of Science Education* (S. 309-323). Dordrecht: Springer.

ERICSSON, K., & SIMON, H. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

GRUBE, C. (2010). Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung.Untersuchung der Struktur und Entwicklungdes wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I. (Dissertation, Universität Kassel, 2010). Verfügbar unter: https://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2011041537247/3/DissertationChristianeGrube.pdf

HAMMANN, M. (2004). Kompetenzentwicklungsmodelle. Merkmale und ihre Bedeutung – dargestellt anhand von Kompetenzen beim Experimentieren. *Mathematischer und Naturwissenschaftlicher Unterricht, 57(4)*, 196-203.

HARTIG, J., & JUDE, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. HARTIG,& E. KLIEME (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (S. 17–30). Bonn: BMBF.

HARTIG, J., & KLIEME, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. SCHWEIZER (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Berlin: Springer.

HARTMANN, S., MATHESIUS, S., STILLER, J., STRAUBE, P., KRÜGER, D., & UPMEIER ZU BELZEN, A. (im Druck). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte: Das Projekt Ko-WADiS. In B. KOCH-PRIEWE, A. KÖKER, J. SEIFRIED, & E. WUTTKE (Hrsg.), *Kompetenzen von Lehramtsstudierenden und angehenden ErzieherInnen*. Bad Heilbrunn: Klinkhardt.

JONKISZ, E., MOOSBRUGGER, H., & BRANDT, H. (2012): Planung und Entwicklung von Tests und Fragebogen. In H. MOOSBRUGGER, & A. KELAVA (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27-74). Heidelberg: Springer.

KELAVA, A., & MOOSBRUGGER, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. MOOSBRUGGER, & A. KELAVA (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 75–102). Heidelberg: Springer.

KLAHR, D. (2000). *Exploring science: The cognition and development of discovery processes.* Cambridge: MIT.

KUNZ, H. (2012). *Professionswissen von Lehrkräften der Naturwissenschaften im Kompetenzbereich Erkenntnisgewinnung* (Dissertation, Universität Kassel, 2012). Verfügbar unter:

https://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2012012040403/9/DissertationHagenKunz.pdf

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.

MARTINEZ, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207-218.

MAYER, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. KRÜGER & H. VOGT (Hrsg.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 177–186). Berlin: Springer.

MESSICK, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

MOOSBRUGGER, H., & KELAVA, A. (2012): Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. MOOSBRUGGER, & A. KELAVA (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7-26). Heidelberg: Springer.

NEUHAUS, B., & BRAUN, E. (2007): Testkonstruktion und Testanalyse – praktische Tipps für empirisch arbeitende Didaktiker und Schulpraktiker. In H. BAYRHUBER, D. ELSTER, D. KRÜGER, & H. J. VOLLMER (Hrsg.): *Forschungen zur Fachdidaktik: Band 9 Kompetenzentwicklung und Assessment* (S. 135-164). Innsbruck: Studien Verlag.

ORSENNE, J. & UPMEIER ZU BELZEN, A. (2012). Hands-On-Aufgaben zur Erfassung und Förderung von Modellkompetenz im Biologieunterricht. In U. HARMS, & F. X. BOGNER (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 5. Didaktik der Biologie – Standortbestimmung und Perspektiven* (S. 29-44), Innsbruck: Studienverlag.

POPPER. K. R. (1934, 2005). *Die Logik der Forschung.* Tübingen: Mohr Siebeck.

SADLER, P. M. (1998). Psychometric models of student conceptions in science: reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35(3),* 265-296.

SCHERER, R., & TIEMANN, R. (2012). Factors of problem-solving competency in a virtual chemistry enviroment: The role of metacognitive knowledge about strategies. *Computers & Education, 59(4),* 1199-1214.

SCHMIEMANN, P., & LÜCKEN, M. (2014). Validität – misst mein Test, was er soll? In D. Krüger, I. Parchmann, & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 107-118). Berlin: Springer-Verlag.

SEKRETARIAT DER STÄNDIGEN KONFERENZ DER KULTUSMINISTER DER LÄNDER IN DER BUNDESREPUBLIK DEUTSCHLAND. [KMK] (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10). Beschluss vom 16.12.2004.* München: Luchterhand.

SEKRETARIAT DER STÄNDIGEN KONFERENZ DER KULTUSMINISTER DER LÄNDER IN DER BUNDESREPUBLIK DEUTSCHLAND. [KMK] (2013). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung*. Berlin: Kultusministerkonferenz. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen _beschluesse/2008/2008_-10_16_Fachprofile-Lehrerbildung.pdf

TERZER, E., PATZKE, C., & UPMEIER ZU BELZEN, A. (2012). Validierung von Multiple-Choice Items zur Modellkompetenz durch lautes Denken. In U. HARMS, & F. X. BOGNER (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 5. Didaktik der Biologie – Standortbestimmung und Perspektiven* (S. 45-62), Innsbruck: Studienverlag.

TERZER, E, HARTIG, J., & UPMEIER ZU BELZEN, A. (2013). Systematische Konstruktion eines Tests zu Modellkompetenz im Biologieunterricht unter Berücksichtigung von Gütekriterien. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 51-76.

UPMEIER ZU BELZEN, A., & KRÜGER, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41-57.

VON AUFSCHNAITER, C., & BLÖMEKE, S. (2010). Professionelle Kompetenz von (angehenden) Lehrkräften erfassen – Desiderata. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 361-367.

WALPUSKI, M. & ROPOHL, M. (2014). Statistische Verfahren für die Analyse des Einflusses von Aufgabenmerkmalen auf die Schwierigkeit. In D. KRÜGER, I. PARCHMANN, & H. SCHECKER (Hrsg.): *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 385-398). Berlin: Springer-Verlag.

WELLNITZ, N., & MAYER, J. (2013). Erkenntnismethoden in der Biologie – Entwicklung und Evaluation eines Kompetenzmodells. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 315-345.