

Wellnitz, Nicole; Mayer, Jürgen

Erkenntnismethoden in der Biologie. Entwicklung und Evaluation eines Kompetenzmodells

Zeitschrift für Didaktik der Naturwissenschaften : ZfDN 19 (2013), S. 315-345



Quellenangabe/ Reference:

Wellnitz, Nicole; Mayer, Jürgen: Erkenntnismethoden in der Biologie. Entwicklung und Evaluation eines Kompetenzmodells - In: Zeitschrift für Didaktik der Naturwissenschaften : ZfDN 19 (2013), S. 315-345 - URN: urn:nbn:de:0111-pedocs-317089 - DOI: 10.25656/01:31708

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-317089>

<https://doi.org/10.25656/01:31708>

in Kooperation mit / in cooperation with:



IPN

Leibniz-Institut für die Pädagogik der
Naturwissenschaften und Mathematik

<https://www.leibniz-ipn.de>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.

This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS

DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Informationszentrum (IZ) Bildung

E-Mail: pedocs@dipf.de

Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

NICOLE WELLNITZ UND JÜRGEN MAYER

Erkenntnismethoden in der Biologie – Entwicklung und Evaluation eines Kompetenzmodells

Scientific methods in biology – development and evaluation of a competence model

ZUSAMMENFASSUNG

Die Biologie nutzt das Beobachten, Vergleichen und Experimentieren als grundlegende Methoden der Erkenntnisgewinnung. Die Beherrschung dieser Methoden und die Auseinandersetzung mit deren Grenzen sind zentrale Aspekte naturwissenschaftlicher Grundbildung. Bisher fehlt ein Modell, welches die methodenspezifischen Kompetenzkonstrukte angemessen abbildet. In diesem Beitrag¹ wird ein entsprechendes Kompetenzmodell präsentiert, dessen Tragfähigkeit mit einem aufgabenbasierten Test überprüft wurde. Die Aufgaben ($N=132$) wurden im Rahmen der Pilotierungsstudie zur „Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I“ mittels eines Facettendesigns konstruiert. Dabei wurden prozessbezogene Teilkompetenzen (*Fragestellung, Hypothese, Untersuchungsdesign* und *Datenauswertung*) vollständig mit methodenbezogenen Teilkompetenzen (*Beobachten, Vergleichen, Experimentieren*) gekreuzt. Zudem wurden diese in fünf Kompetenzniveaus graduert. Die Daten von 986 Schülerinnen und Schülern zeigen, dass die zur inhaltlichen Charakterisierung herangezogenen Aufgabenmerkmale erwartungsgemäß mit den Aufgabenschwierigkeiten korrespondieren.

Schlagwörter: Beobachten, Erkenntnisgewinnung, Experimentieren, Kompetenzmodell, Ordnen, Vergleichen

ABSTRACT

In order to create an environment that supports students' understanding of different inquiry methods, it is important to analyse how methodological competence can be described, differentiated and assessed. This article presents a competence model for *observing, comparing* and *experimenting*. Each method is divided into the process-related skills *question, hypothesis, research design* and *data analysis*. To specify different requirements, each particular facet of inquiry method and process skill is graduated hierarchically in five competence levels. This structure was used to develop an instrument based on 132 multiple choice, short and open-ended items. To evaluate the quality of the test,

1 Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 GJ 0862 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

data of 986 students participating in the German project “Evaluation of the National Educational Standards for Natural Sciences at the Lower Secondary Level” were analysed. The findings provide evidence for the correlation between content-related task characteristics and task difficulty.

Keywords: classifying, comparing, competence model, experimentation, observation, scientific inquiry

1 Einleitung

Lernende und teils auch Lehrende gehen vielfach davon aus, dass Wissenschaftler auf der Suche nach Erklärungen für naturwissenschaftliche Phänomene eine einzige universelle wissenschaftliche Methode heranziehen (Abd-El-Khalick & Lederman, 2000; McComas, 1996; Windschitl, 2004). Unter *einer wissenschaftlichen Methode* wird dabei eine rezeptartige Vorgehensweise mit algorithmisch abzuarbeitenden Schritten verstanden, die unzweifelhafte Ergebnisse hervorbringen (Bell, Blair, Crawford & Lederman, 2003). Diese Vorstellung wird durch den Status, den experimentelle Herangehensweisen in Lehrplänen, Unterrichtsmaterialien und Schulbüchern besitzen, erhärtet (Lederman, 2004). Zum einen wird der Prozess der naturwissenschaftlichen Erkenntnisgewinnung oftmals am Beispiel des Experimentierens dargestellt. Zum anderen wird der Terminus Experiment zur sprachlichen Vereinfachung übergeneralisierend für jegliche Art von Schulversuchen verwendet. Diese recht großzügige Auslegung des Begriffs *Experiment* geht soweit, dass darunter sowohl klassische Nachweisreaktionen als auch der generelle Umgang mit

wissenschaftlichen Instrumenten (Mikroskopieren, Präparieren) verstanden wird (Wellington, 2002). Tatsächlich, zumindest in wissenschaftsmethodischer Hinsicht, handelt es sich dabei um wissenschaftliche Arbeitstechniken, die vorrangig manuelle Fertigkeiten erfordern oder um naturwissenschaftliche Untersuchungen (Beobachtungen, Vergleiche, Experimente), die ein hypothetisch-deduktives Vorgehen zur Beantwortung naturwissenschaftlicher Fragen verlangen. Ähnliche terminologische Unschärfen sind bei der Verwendung der Begriffe *Beobachten* und *Vergleichen* zu konstatieren. Zu Beginn und während experimenteller Forschungsprozesse werden Phänomene beobachtet, Beobachtungsdaten beschrieben und miteinander verglichen. Beobachtungen und Vergleiche werden dementsprechend als Vorstufen oder Teilaspekte des Experimentierens verstanden (Darian, 1997), die letztendlich im Experiment, dem ‚Königsweg der Erkenntnisgewinnung‘, münden. Dabei bleibt unberücksichtigt, dass die experimentelle Methode nicht auf alle naturwissenschaftlichen Probleme anwendbar ist. Die Entdeckung von Floren und Faunen sowie ihre Systematisierung, Ökologie und Ethologie

sind nur einige Beispiele, die vielfach auf Beobachtungen und Vergleiche zurückzuführen sind (Mayr, 1984).

In den nationalen Bildungsstandards Biologie des Kompetenzbereichs *Erkenntnisgewinnung* wird konsequenterweise neben der Anwendung von Schritten aus dem experimentellen Weg auch explizit das Verständnis und der Umgang mit weiteren charakteristischen Methoden wie dem Beobachten und dem Vergleichen als Bildungsanspruch formuliert (KMK, 2005). Für ein elaboriertes Verständnis naturwissenschaftlicher Erkenntnisgewinnung ist diese Differenzierung aus fachdidaktischer Sicht in mehrfacher Hinsicht notwendig:

Schülerinnen und Schüler sollen verstehen,

- dass naturwissenschaftliche Erkenntnisse nicht mittels *der wissenschaftlichen Methode*, sondern mit verschiedenen Methoden generiert werden,
- dass beim Beobachten Merkmalszusammenhänge ohne Manipulation von Variablen untersucht werden,
- dass beim Vergleichen trennscharfe Kriterien zur Bildung von Ähnlichkeits- oder Verwandtschaftsgruppen herangezogen werden,
- dass beim Experimentieren eine Einflussgröße zur Prüfung von Ursache-Wirkungs-Beziehungen mittels Kontroll- und Experimentalansatz variiert wird,
- dass Zusammenhangssysteme (Korrelation), Ordnungssysteme sowie Ursache-Wirkungssysteme (Kausalität) unterschiedliche Differenzierungen naturwissenschaftlicher Erkenntnismethoden kennzeichnen.

Bisher fehlt ein theoretisch hergeleitetes und empirisch geprüftes Modell mit dem die wissenschaftsmethodischen Kompetenzen von Schülerinnen und Schülern der Sekundarstufe I methodenspezifisch beschrieben, gefördert und letztendlich auch erfasst werden können.

2 Wege der Erkenntnis

In den Naturwissenschaften wird der Prozess zur Aufstellung und Begründung von Hypothesen und Theorien als forschungslogischer Ablauf mit verschiedenen Teilschritten beschrieben. Grundlage ist zumeist das Wissenschaftsmodell von Popper (1994), welches naturwissenschaftliche Erkenntnisgewinnung als hypothetisch-deduktive Vorgehensweise modelliert. Der auf diesem Modell basierende Erkenntnisgewinn wird durch Verwerfung sich nicht bewährender Theorien mittels empirischer Falsifikation erlangt. In der Fachdidaktik werden die Teilschritte zur erfolgreichen Generierung naturwissenschaftlicher Erkenntnisse mit verschiedenen, zu erwerbenden Teilkompetenzen systematisch in Beziehung gesetzt. Je nach zugrundeliegender Konzeption werden die zu fördernden Kompetenzen im Hinblick auf Anzahl und Details in unterschiedlichem Maße ausdifferenziert. Übereinstimmend wird sowohl in empirischen Studien als auch in nationalen und internationalen Standards und Curricula (für eine detaillierte Übersicht der Strukturierung siehe Emden & Sumfleth, 2012, S. 69) die Bedeutung der Teilkompetenzen Fragestellungen und Hypothe-

sen formulieren, Untersuchungen planen und durchführen sowie Daten auswerten und interpretieren herausgestellt (z. B. DfES & QCA, 2004; Klahr, 2000; KMK, 2005; Mayer, Grube & Möller, 2008). Während in den deutschen Bildungsstandards nicht streng zwischen der Formulierung von naturwissenschaftlichen Fragen und Hypothesen, der Planung und Durchführung sowie der Auswertung und Interpretation von Daten unterschieden wird, sind die amerikanischen Standards weit aus differenzierter. Es werden acht sogenannte Praktiken beschrieben, die im naturwissenschaftlichen Unterricht gefördert werden sollen: „asking questions“, „developing and using models“, „planning and carrying out investigations“, „analyzing and interpreting data“, „using mathematics and computational thinking“, „constructing explanations“, „engaging in argument from evidence“ und „obtaining, evaluating, and communicating information“ (NRC, 2011, S. 3).

Den meisten Ansätzen ist gemein, dass naturwissenschaftliche Erkenntnisgewinnung nicht als stringenter, sondern als rekursiver und iterativer hypothetisch-deduktiver Prozess verstanden wird, der auf vielgestaltige Art und Weise erfolgen kann (Harwood, 2004). Aus lerntheoretischer Perspektive wird der Erkenntnisprozess als komplexer Problemlöseprozess aufgefasst, der durch grundlegende Prozeduren charakterisiert und von Situations- und Personenvariablen beeinflusst wird (Doran, Helgeson & Kumar, 1995; Klahr, 2000; Mayer, 2007). Je nach qualitativer Differenz zwischen einer (naturwissenschaftlichen) Fragestellung als Ausgangs-

zustand und einer potentiellen Antwort als Zielzustand sind unterschiedliche kognitive Aktivitäten zur Erreichung des Zielzustandes erforderlich (Klix, 1996). Mit anderen Worten: Je nach Problemstellung gilt es, den entsprechenden Erkenntnisweg bzw. die passende Methode auszuwählen und deren Vor- und Nachteile sowie ihre Grenzen zur Prüfung von Hypothesen und Theorien abzuwägen.

2.1 Beobachten

Der Begriff *Beobachten* wird im Alltag und in der Wissenschaft mit verschiedenen Bedeutungen verwendet. Im Alltag werden, in Abgrenzung zur spontanen Wahrnehmung (z. B. Ich nehme einen Vogelschwarm wahr), alle selektiven Wahrnehmungen, die mit einer gezielten Suche verbunden sind (z. B. Ich halte Ausschau nach einem bestimmten Vogel), als Beobachtungen bezeichnet (Mahner & Bunge, 2000). Gegenstand der Beobachtung sind nur sinnlich (direkte Beobachtung) oder apparativ (indirekte Beobachtung)² zu erfassende statische oder dynamische As-

2 In der fachdidaktischen Literatur werden indirekte, d. h. instrumentell vermittelte Beobachtungen auch als Untersuchungen bezeichnet (Gropengießer, 2013; Killermann, Hiering & Starosta, 2005; Köhler, 2004). Die Nutzung von Hilfsmitteln erlaubt die quantitative Erweiterung der Sinnesorgane (z. B. mittels Lichtmikroskop oder Waage) oder konvertiert bestimmte Informationen derart, dass sie in anderer Form sichtbar werden (z. B. mittels Elektronenmikroskop oder Nachweismittel) (Flannery, 2007; Mahner & Bunge, 2000). In diesem Beitrag wird der Terminus *Untersuchung* als Oberbegriff für alle hypothetisch-deduktiven Vorgehensweisen verwendet.

pekte beispielsweise biologischer Systeme (Zellen, Organismen, Ökosysteme etc.), die weder sprachlich vermittelt werden noch auf Dokumenten beruhen. Alltagsbeobachtungen finden eher zufällig statt. Sie sind von Spontaneität und Subjektivität geleitet und obliegen keinem genauen Beobachtungsplan (Bortz & Döring, 2006). In der Wissenschaft wird unter *Beobachtung* eine systematische Vorgehensweise verstanden, die geplant und intersubjektiv prüfbar ist (Bortz & Döring, 2006). Systematische Beobachtung findet dabei auf zwei Ebenen statt: Zum einen ist sie Teil verschiedener Untersuchungen (Beobachtung, Vergleich, Experiment), zum anderen kennzeichnet sie eine eigene wissenschaftliche Methode, die ein spezifisches Beobachtungsdesign erfordert. Ersteres bedeutet, dass beispielsweise beim Experimentieren zur objektiven Beschreibung der experimentell gewonnenen Daten, die beobachtbaren Endzustände des Messprozesses bzw. der Sinnesdaten systematisch beobachtet und erfasst werden (Graßhoff, Casties & Nickelsen, 2000). Um dem Kriterium der Objektivität zu genügen, darf die konkrete Beschreibung von Sinnesdaten nicht mit der Interpretation des Beschriebenen einhergehen (Norris, 1984).

Davon abzugrenzen ist die systematische Beobachtung als gezielte wissenschaftliche Methode der Datenerhebung zur Beantwortung expliziter Fragestellungen (z. B. Wie bewegen sich Wasserflöhe fort?, Tab. 1) und zur Prüfung von Hypothesen (Bortz & Döring, 2006; Greve & Wentura, 1997). Ziel ist die Beschreibung korrelativer oder funktionaler Merkmalszusam-

menhänge statischer oder dynamischer Systemeigenschaften, ohne dass zuvor aktiv Variablen manipuliert wurden. Beobachtungen zu korrelativen Zusammenhängen hinterfragen Beziehungen zwischen zwei oder mehreren Merkmalen³ eines biologischen Systems (z. B. den Zusammenhang zwischen Fortbewegung und morphologischen Merkmalen, Tab. 1). Beobachtungen zu funktionalen Zusammenhängen dienen der Erfassung von Form- und Funktionsbeziehungen, indem beispielsweise die Funktion von ausgewählten Merkmalen eines Systems ermittelt wird (Boerwinkel, Waarlo & Boersma, 2009). Dabei muss gezielt zwischen beobachtungsrelevanten und unwesentlichen Merkmalen differenziert werden (Selektion). Je nach Fragestellung werden diese Merkmale qualitativ oder quantitativ unter Kontrolle von Störvariablen gemessen und dokumentiert. Bei der Deutung der Ergebnisse dürfen korrelative oder funktionale Zusammenhänge nicht als Kausalzusammenhänge interpretiert werden. Die biologische Funktion eines Merkmals sagt beispielsweise noch nichts über ihre physiologische Verursachung aus (Carrier, 2005).

3 Als Merkmale werden in Anlehnung an Klix (1996) die „für einen Gegenstand klassifizierungsrelevanten Objekteigenschaften“ (Klix, 1996, S. 536) bezeichnet. Nach der Logik sind es die Merkmale, wodurch sich Objekte voneinander unterscheiden. „Es können entweder Eigenschaften oder Beziehungen sein“ (Šula, 1968, S. 22).

2.2 Vergleichen

Beim Vergleichen als Methode naturwissenschaftlicher Erkenntnisgewinnung werden biologische Systeme (z. B. Organismen, Tab. 1) auf Gemeinsamkeiten und Unterschiede geprüft. Dies erfolgt mit dem Ziel, Ordnungssysteme zu generieren, die Ähnlichkeits- oder Verwandtschaftsgruppen abbilden. Aus Sicht der Logik wird das Vergleichen als dreistellige Relation beschrieben, bei der mindestens zwei Vergleichsobjekte durch Festlegung eines übergeordneten Kriteriums zueinander in Beziehung gesetzt werden (Janich & Weingarten, 1999). Erkenntnistheoretisch vollzieht sich der Vergleich in vier Schritten: (a) Durch ein Problem, eine Fragestellung oder eine Hypothese wird ein Vergleich initiiert. Die Frage nach Unterschieden und / oder Gemeinsamkeiten soll geklärt werden. (b) Aus der Analyse der Vergleichsobjekte werden zweckgerichtet Vergleichskriterien abgeleitet und kategorisiert. (c) Die Vergleichsobjekte werden herangezogen und der Vergleich wird durchgeführt. (d) Abschließend werden erfasste Unterschiede und Gemeinsamkeiten dokumentiert und bewertet (Eichberg, 1972). Neben den Vergleichskriterien (z. B. Anzahl der Beinpaare) müssen deren Ausprägungen (z. B. 6, 10, 14) berücksichtigt werden, indem bei der Bildung von Gruppen eines Klassifikationssystems nur die Ausprägungen eines Kriteriums herangezogen werden. Nur diese Ordnungssysteme werden als kriterienstet bezeichnet (Hamman, 2002). Je mehr Kriterien einem Vergleich zugrunde liegen (z. B. Anzahl der Beinpaare und Körpergliederung), desto elaborierter sind

die ermittelten Ordnungssysteme (Hamman, 2005). Über den Vergleich ist es z. B. in der Biologie möglich, Erklärungen für phylogenetisch, ökologisch oder allgemeinbiologisch bedingte Ähnlichkeiten zu erhalten.

2.3 Experimentieren

Beim Experimentieren werden in der Regel kausale Zusammenhänge (= Ursache-Wirkungs-Beziehungen) zwischen Variablen untersucht (Schulz & Wirtz, 2012; Schulz, Wirtz & Starauschek, 2012). Naturwissenschaftliche Phänomene und Probleme werden auf Ursachen zurückgeführt und damit erklärt (Graßhoff et al., 2000). Im Gegensatz zum Beobachten als Erkenntnismethode werden beim Experimentieren Messungen an einem vom Forscher kontrolliert veränderten Erkenntnisgegenstand vorgenommen (Mahner & Bunge, 2000). Die „invasive Tendenz und die Kontrolle, die der Experimentator im Experiment über den Gegenstand und dessen Umgebung gewinnt“ (Köchy, 2006, S. 79) zeichnen diese Methode aus. Drei Variablen sind hierbei von Bedeutung: die unabhängige Variable (Einflussgröße), die abhängige Variable (Messgröße) und die Kontrollvariable (Störgröße). Die unabhängige Variable (z. B. Licht) wird gezielt verändert, um deren Auswirkung auf die abhängige Variable (z. B. Bewegungsreaktion) zu bestimmen (Tab. 1).

Um einen direkten Effekt der abhängigen Variablen beobachten und messen zu können, müssen alle anderen Variablen, die ebenfalls einen Einfluss auf die abhängige

Tab. 1: Beobachten, Vergleichen und Experimentieren am Beispiel des Wasserfloh (Meier & Wellnitz, 2013; verändert)

	BEOBSACHTEN	VERGLEICHEN	EXPERIMENTIEREN
Frage	Wie bewegen sich Wasserflöhe fort?	Ist der Wasserfloh ein Floh?	Reagieren Wasserflöhe auf Licht?
Hypothese	H₁: Wasserflöhe schweben passiv im Wasser. H₂: Wasserflöhe gleiten aktiv durch das Wasser. H₃: Wasserflöhe hüpfen aktiv durch das Wasser.	H₀: Wasserflöhe und Hundeflöhe gehören verschiedenen taxonomischen Gruppen an. H₁: Wasserflöhe und Hundeflöhe gehören zur Gruppe der Insekten.	H₀: Wasserflöhe zeigen keine Reaktion auf Licht. H₁: Wasserflöhe schwimmen zum Licht. H₂: Wasserflöhe schwimmen vom Licht weg.
Untersuchungsdesign	Richtung der Fortbewegung und Bewegung der Körperanhänge werden in einem durchsichtigen Behälter bei Tageslicht beobachtet und dokumentiert.	Verschiedene Organismen, z. B. Wasser-, Hundefloh, Libellenlarve und Wasserasel, werden kriteriengeleitet (z. B. Anzahl der Beinpaare) verglichen und in Gruppen eingeteilt.	Ein halbseitig verdunkeltes Becken mit Wasserflöhen wird mit Licht bestrahlt. Der Aufenthaltsort der Wasserflöhe wird dokumentiert.
Datenauswertung	Zwei Fortbewegungsphasen sind zu beobachten: Ein Schweben mit gleichzeitigem Absinken und eine hüpfende, flohartige Bewegung. Wasserflöhe bewegen sich durch ruderartiges Schlagen mit dem zweiten Antennenpaar fort. Das Absinken mit ausgebreiteten Antennen in der Schwebephase begünstigt die hüpfende Bewegung.	Entwicklung eines Ordnungssystems mit zwei Gruppen: Insekten mit 3 Beinpaaren (Hundefloh, Libellenlarven) und Krebstiere mit 5 oder mehr Beinpaaren (Wasserfloh, Wasserasel). Wasser- und Hundeflöhe gehören trotz ähnlicher Namen zu unterschiedlichen taxonomischen Gruppen.	Wasserflöhe bevorzugen den äußeren, weniger hellen Bereich einer Lichtquelle. Wasserflöhe orientieren sich zum Licht, wenn es nicht zu stark (hell) ist. Es handelt sich um einen Faktor, an dem sie sich in den Wasserschichten orientieren können.

Variable haben könnten (z. B. Temperatur), gezielt konstant gehalten werden. Dafür sind mindestens ein Experimentalansatz (z. B. Wasserfloh im Licht) und ein Kontrollansatz (z. B. Wasserfloh ohne Licht) erforderlich. Dies geschieht immer unter der Voraussetzung, dass sich die unabhängige Variable manipulieren und Störgrößen kontrollieren lassen. Bei Fragestellungen

zu qualitativen Kausalprozessen wird überprüft, inwiefern ein bestimmter Faktor überhaupt in einem kausalen Zusammenhang zu einem anderen bestimmten Faktor steht. Quantitative Kausalprozesse betreffen hingegen die Frage, wie sich verschiedene Werte eines Faktors auf die Werte eines anderen Faktors auswirken (Graßhoff et al., 2000).

3 Empirische Befunde

3.1 Lernvoraussetzungen beim Beobachten

Trotz der Bedeutung naturwissenschaftlicher Beobachtungen liegen nur wenige Studien vor, die diesbezüglich Stärken und Schwächen von Schülerinnen und Schülern untersuchen. Ein Kompetenzentwicklungsmodell zum Beobachten im Vor- und Elementarschulbereich wurde von Kohlhauf, Rutke und Neuhaus (2011) entwickelt und evaluiert. Als Teilkompetenzen biologischer Beobachtungskompetenz konnten „Beschreiben“, „Fragen / Vermuten / Testen“ und „Interpretieren“ als unabhängige Skalen mit jeweils drei Ausprägungen („Inzidentelles Beobachten“, „Unsystematisches Beobachten“, „Systematisches Beobachten“) empirisch identifiziert werden (Kohlhauf et al., S. 219). Mit zunehmendem Ausbildungsstand (vom Kindergarten bis zum Studium) steigt die Beobachtungskompetenz erwartungsgemäß signifikant an (Kohlhauf, 2012). Tomkins und Tunnicliffe (2001) fanden heraus, dass zwölfjährige Schülerinnen und Schüler bei der Kurzzeitbeobachtung von Salinenkrebsen vorrangig hervorstechende anatomische Merkmale oder Verhaltensweisen wahrnehmen, ohne diese zu hinterfragen. Erst bei Langzeitbeobachtungen der Krebse wurden von den Lernenden zu überprüfende Hypothesen formuliert. Wenn Lernende der 9. Jahrgangsstufe zu einem gegebenen naturwissenschaftlichen Problem, welches ausschließlich durch die Methode des Beobachtens gelöst werden kann, Fragestellungen oder

Hypothesen formulieren sollen, geschieht dies, wie beim Experimentieren und Vergleichen, unter Missachtung des Fachwissens (Arnold, Wellnitz & Mayer, 2011). Für höhere Jahrgänge konnten Aydeniz, Baksa und Skinner (2011) zeigen, dass ein angemessenes Verständnis des Unterschieds zwischen einem Experiment und einer Beobachtung vorhanden ist. Für die Mehrheit von Lehramtsstudenten sind Beobachtungen allerdings „just noticing things“ (Ahtee, Suomela, Juuti, Lampiselkä & Lavonen, 2009, S. 138). Untersuchungen von Beobachtungskompetenzen bei Grundschulreferendaren zeigen, dass angehende Lehrkräfte beim Beobachten von biologischen Systemen nicht systematisch vorgehen und keine wissenschaftlichen Kriterien einhalten (Oğuz & Yurumezoğlu, 2007; 2009). Beobachtungen werden mehr oder weniger wahllos durchgeführt und ähneln Beobachtungen wie sie im Alltag vonstatten gehen. Es werden keine Forschungsfragen formuliert und keine Erkenntnisobjekte mit zu messenden Merkmalen festgelegt. Stattdessen werden mehrere biologische Systeme gleichzeitig beobachtet, z. B. Insekten auf Laubblättern, ohne einen Zusammenhang zwischen beiden Systemen herzustellen. Nur wenn strukturierte Beobachtungsbögen zur Verfügung gestellt werden, findet eine mehr oder weniger systematische Vorgehensweise statt. Es werden Forschungsfragen entwickelt, die jedoch oftmals nicht den wissenschaftlichen Anforderungen genügen (Oğuz & Yurumezoğlu, 2007).

3.2 Lernvoraussetzungen beim Vergleichen

In älteren Studien, welche die Fähigkeiten von Lernenden beim Vergleichen untersuchen, steht die Erfassung taxonomischen Wissens im Vordergrund und weniger die methodische Herangehensweise. So wird zumeist geprüft, inwiefern systematische Kriterien bei der Erstellung taxonomischer Gruppen berücksichtigt werden. Beispielsweise sollen von den Schülerinnen und Schülern ausgewählte Organismen vorgegebenen taxonomischen Gruppen zugeordnet werden (Braund, 1991; Ryman, 1974; Trowbridge & Mintzes, 1988). Das Augenmerk aktueller Studien liegt dagegen eher auf der Vorgehensweise bei der Erstellung von Ordnungssystemen oder hierarchischen Systemen, d. h. wie und mit welchen Kriterien biologische Systeme gruppiert und ggf. systematisiert werden. Zumeist werden Schülerinnen und Schüler aufgefordert, vorgegebene Pflanzen oder Tiere in Gruppen frei einzuteilen und diesen Gruppen beschreibende Namen zu geben. Alternativ sollen vorgegebene Gruppen unter Angabe einer Begründung um weitere Organismen ergänzt oder einzelne Organismen ausgesondert werden (Hammann, 2002; Kattmann & Schmitt, 1996; Krüger & Burmester, 2005; Sonnefeld & Kattmann, 2002).

Die Ergebnisse zeigen, dass es den Lernenden beim Vergleichen und Ordnen von Organismen schwer fällt, Kriterienwechsel zu vermeiden. Es werden für die Bildung mehrerer Gruppen eines Ordnungssystems die Ausprägungen verschiedener, nicht trennscharfer Kriterien herangezo-

gen, indem beispielsweise gleichzeitig nach beschreibenden (z. B. länglich, rund), ökologischen (z. B. Raubfisch, Pflanzenfresser) und verwandtschaftlichen Kriterien (z. B. Haie, Karpfen) sortiert wird (Hammann, 2002; Hammann & Bayrhuber, 2003). Diese von Hammann (2004b) als kriterienuntest bezeichneten Klassifikationssysteme ähneln den Ordnungssystemen, die innerhalb des Alltags Anwendung finden. Mit höherer Jahrgangsstufe nimmt zudem die Wahrscheinlichkeit für eine kriterienstetige Ordnung ab. Dies ist darin begründet, dass mit zunehmendem Alter mehr Gruppen gebildet werden. Grundschüler bilden beim Ordnen von Objekten unter Berücksichtigung eines übergeordneten Kriteriums (kriterienstet) Ordnungssysteme mit durchschnittlich zwei Gruppen (George & Dietz, 1971). Dagegen legen Sekundarschüler innerhalb eines Ordnungssystems drei bis fünf Gruppen an, die nicht immer trennscharf zueinander sind (Krüger & Burmester, 2005). Biologisch-systematische Kriterien sind beim Ordnen von Organismen generell unterrepräsentiert (Hammann, 2002; Krüger & Burmester, 2005; Sonnefeld & Kattmann, 2002).

3.3 Lernvoraussetzungen beim Experimentieren

Nationale und internationale Studien belegen deutliche Schwächen von Lernenden bei der Bearbeitung von schriftlichen oder praktischen Experimentieraufgaben. Die Schülerinnen und Schüler haben Schwierigkeiten aus einem vorgegebenen

naturwissenschaftlichen Phänomen Fragestellungen abzuleiten, die mit einem experimentellen Design beantwortet werden können (Mayer et al., 2008). Hypothesen zu dargebotenen experimentellen Designs werden nicht passend ausgewählt (Hammann, Phan, Ehmer & Bayrhuber, 2006) oder Experimente gänzlich ohne Hypothesen durchgeführt (Hammann, 2004a; Meier & Mayer, 2012; Möller, Hartmann & Mayer, 2010).

Bei der Durchführung von Experimenten wird häufig lediglich ein Ansatz getestet. (Hammann, Phan, Ehmer & Grimm, 2008; Möller et al., 2010). Dabei wird nicht berücksichtigt, dass nur durch den Vergleich eines Experimentalansatzes mit einem Kontrollansatz Ursache-Wirkungs-Zusammenhänge ermittelt werden können. Mit zunehmender Komplexität eines experimentellen Designs, z. B. wenn die Auswirkungen von mehreren unabhängigen Variablen auf eine zu messende Variable untersucht werden sollen, fällt es Schülerinnen und Schülern schwer, die relevanten Variablen zu identifizieren und gezielt zu variieren (Duggan, Johnson & Gott, 1996). Auch wenn ein Kontrollansatz berücksichtigt wird, werden oftmals mehrere Variablen gleichzeitig verändert, ohne diese konsequent zu isolieren und zu kontrollieren (Arnold, Kremer & Mayer, 2012). Diese sogenannten konfundierten Experimente lassen keine eindeutigen Aussagen über die Wirkung einzelner unabhängiger Variablen zu. Zusätzliche Schwierigkeiten bereitet die Operationalisierung kontinuierlicher Variablen. Es werden auch dann kategoriale Variablen herangezogen, wenn diese

für die Überprüfung eines Zusammenhanges ungeeignet sind (Duggan et al., 1996). Experimente werden auch so geplant, dass nur die Variablen verändert werden, von denen Schülerinnen und Schüler annehmen, dass diese ihre eigenen Überzeugungen bestätigen (Klahr, Fay & Dunbar, 1993). Lernende bemühen sich nicht um Belege, die ihre Hypothese widerlegen könnten.

Bei der Auswertung experimenteller Daten stellen Schülerinnen und Schüler z. T. keinen Bezug zur Hypothese her. Beobachtete Effekte werden lediglich beschrieben, ohne deren Ursachen zu erklären (Germann, Aram & Burke, 1996). Ohne Berücksichtigung eines Kontrollansatzes wird auf Kausalität geschlossen (Hammann, Phan & Bayrhuber, 2007). Die Lernenden halten selbst bei Daten, die ihre Hypothesen widerlegen, weiter an den Ausgangshypothesen fest. Bestätigenden Befunden wird eine größere Bedeutung beigemessen, während unerwartete Ergebnisse ignoriert, umgedeutet oder auf methodische Fehler zurückgeführt werden (Chinn & Brewer, 1998).

4 Theoriegeleitete Entwicklung eines Kompetenzmodells zum Beobachten, Vergleichen und Experimentieren

Viele Studien zu experimentellen Kompetenzen legen das von Klahr (2000) postulierte Scientific Discovery as Dual Search Model (SDDS-Modell) mit den drei Teilkompetenzen *Suche im Hypothesenraum*, *Testen von Hypothesen* und *Analyse von*

Evidenzen zugrunde (z.B. Hammann et al., 2007; Klos, Henke, Kieren, Walpuski & Sumfleth, 2008). Mayer (2007) erweitert dieses Modell und benennt in dem Strukturmodell zum Wissenschaftlichen Denken u.a. als zentrale Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung vier Prozessvariablen, die als eigenständige Teilkompetenzen beim Experimentieren empirisch bestätigt werden konnten: naturwissenschaftliche Fragen formulieren, Hypothesen generieren, Untersuchungen planen und Ergebnisse deuten (Grube, 2011; Mayer et al., 2008). Für jede Prozessvariable werden fünf Kompetenzniveaus ausdifferenziert, die sich hinsichtlich Komplexität und qualitativer Abstufung im Sinne eines problemlösenden Erkenntnisprozesses unterscheiden. Die Annahme einer hierarchisch ansteigenden Schwierigkeit konnte für drei – *Untersuchung eines Faktors* (Niveau I), *Untersuchung von Zusammenhängen* (Niveau II) und *Kontrollierte Untersuchung auf Basis von Konzeptverständnis* (Niveau III) – von fünf beschriebenen Kompetenzniveaus belegt werden (Möller et al., 2010). Die von Mayer et al. (2008) identifizierten Prozessvariablen sind als zu überprüfende Aspekte des Bereichs *Erkenntnisgewinnung* unter der Benennung *Fragestellung*, *Hypothese*, *Untersuchungsdesign* und *Datenauswertung* in das fachübergreifende Kompetenzmodell zur „Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I“ (ESNaS) eingeflossen (Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010; Walpuski et al., 2010).

4.1 Ausdifferenzierung methodenspezifischer Teilkompetenzen (Kompetenzstruktur)

Das prozessbezogene Kompetenzkonstrukt *Fragestellung*, *Hypothese*, *Untersuchungsdesign* und *Datenauswertung* wird in dieser Studie herangezogen und theoriegeleitet für die drei Erkenntnismethoden *Beobachten*, *Vergleichen* und *Experimentieren* ausdifferenziert, d.h. um methodenspezifische Kompetenzkonstrukte erweitert. Dabei werden unter dem Terminus *Erkenntnismethoden* explanative, d.h. hypothesenprüfende Untersuchungen verstanden, die das spezifisch hypothetisch-deduktive Vorgehen der Naturwissenschaften verlangen (Wellnitz, 2012). Die in den Bildungsstandards ausgewiesenen Kompetenzerwartungen hinsichtlich des hypothetisch-deduktiven Erkenntnisprozesses einerseits sowie der drei Untersuchungsmethoden (*Beobachten*, *Vergleichen*, *Experimentieren*) andererseits werden in einem Facettendesigns abgebildet, bei dem prozessbezogene mit methodenspezifischen Kompetenzen vollständig gekreuzt werden (Tab. 2).

Je nach Ausgangssituation (Problem, Phänomen) ist eine naturwissenschaftliche Fragestellung zu korrelativen bzw. funktionalen Zusammenhängen (*Beobachten*), zu Ähnlichkeiten bzw. Unterschieden (*Vergleichen*) oder kausalen Zusammenhängen (*Experimentieren*) innerhalb eines Biosystems oder zwischen mehreren Biosystemen zu generieren. Die Konkretisierung der Fragestellung in Form von Hypothesen erfordert die konkrete Nennung der in Beziehung zu setzenden

Tab. 2: Kompetenzstruktur naturwissenschaftlicher Erkenntnismethoden (Wellnitz, 2012; Wellnitz & Mayer, 2012)

	ERKENNTNISMETHODEN		
	Beobachten	Vergleichen	Experimentieren
Fragestellung	Welche Zusammenhänge bestehen zwischen Systemen und ihren Merkmalen sowie deren räumlichen und zeitlichen Auftreten?	Welche Unterschiede oder Gemeinsamkeiten weisen verschiedene Systeme auf? Welche Systeme lassen sich einer Kategorie zuordnen?	Welche Ursache liegt einem Phänomen oder Problem zu Grunde?
Hypothese	System A besitzt Merkmal X mit der Funktion Y. System A und B weisen die räumliche Beziehung AB auf.	Die Systeme A und B gleichen sich im Kriterium X und unterscheiden sich im Kriterium Y.	Die unabhängige Variable (UV) hat einen Einfluss auf die abhängige Variable (AV).
Untersuchungsdesign	Planung, Durchführung und Datengewinnung • <u>Stichprobe</u>: Wahl des Erkenntnisobjektes, inkl. Anzahl, ggf. Randomisierung • <u>Design</u>: Auswahl und Einsatz der technischen Hilfsmittel sowie Versuchsaufbau • <u>Messkonzept</u>: qualitativ/quantitativ, Größen, Messbereich, Skalierung, Messgeräte • <u>Messprotokoll</u>: Anzahl der Messungen, Versuchswiederholungen, Messzeitpunkte • <u>Dokumentation der Rohdatensätze</u>: Beschreibung, Zeichnung, Messwerte, Fotos etc.		
	Beobachtung eines Systems • Identifizierung der beobachtungsrelevanten Merkmale • Identifizierung von Störgrößen (unwesentliche Merkmale) • Selektion zu messender Merkmale • Messung statischer oder dynamischer Systemeigenschaften • Kontrolle von Störgrößen	Vergleich von mindestens zwei Systemen • Identifizierung der Vergleichskriterien • Identifizierung von Störgrößen (Merkmalsvariation) • Auswahl mehrerer übergeordneter (trennscharfer) Kriterien • Berücksichtigung mehrerer Ausprägungen eines Kriteriums • Innergruppen- oder Zwischengruppenvergleich	Manipulation eines Systems • Identifizierung der Messgröße (AV) und Einflussgröße (UV) • Identifizierung von Störgrößen (Kontrollvariable) • systematische Variation der Einflussgröße • Messung der abhängigen Variable • Konstanthaltung der Störgrößen • Kontroll- und Experimentalansatz
	Zusammenhangssystem: Beschreibung der Merkmale, Strukturen und Veränderungen von Systemen	Ordnungssystem: Matrix von Systemen mit übergeordnetem Vergleichskriterium	Ursache-Wirkungssystem: Wirkungsaussage von unabhängiger auf abhängige Variable
Datenauswertung	Daten analysieren und Schlussfolgerungen ziehen • <u>Datenanalyse</u>: Berechnung (z. B. Mittelwert, Steigung) und Darstellung (z. B. Tabelle, Diagramm) • <u>Überprüfung</u> der Datenqualität: Genauigkeit, Fehler, Störgrößen, Methodendiskussion • <u>Rückbindung</u>: Vergleich der Ergebnisse mit der Hypothese (Falsifizierung/vorläufige Bestätigung) • <u>Interpretation</u>: hypothesenbezogene Deutung der Ergebnisse		
	Auswertung korrelativer und funktionaler Zusammenhänge Beschreibung der Zusammenhänge durch... • Form-Funktions-Beziehungen • zeitliche Abfolgen • räumliche Beziehungen	Auswertung von Gruppen und hierarchischen Systemen Beschreibung der Ordnung durch... • phylogenetisch, • ökologisch oder • allgemeinbiologisch bedingte Ähnlichkeiten	Auswertung kausaler Zusammenhänge Erklärung der Ursache eines Phänomens durch... • physiologische, • ökologische oder • genetische Mechanismen

Merkmale (Beobachten), Kriterien (Vergleichen) oder Variablen (Experimentieren). Bei der Planung der Untersuchung wird festgelegt, was beim Beobachten, Vergleichen oder Experimentieren genau betrachtet bzw. beobachtet, beschrieben und gemessen werden soll. Der Erkenntnisgegenstand wird unter Beachtung der zu untersuchenden qualitativen oder quantitativen Merkmale (Beobachten), Vergleichskriterien (Vergleichen) bzw. Variablen (Experimentieren) ausgewählt und entsprechende Hilfsmittel, mit denen das Phänomen erfasst und gemessen werden kann, herangezogen. Während beim Beobachten keine auf das Erkenntnisobjekt beeinflussenden Veränderungen vorgenommen werden, die sich auf die zu messenden Merkmale statischer oder dynamischer Systemeigenschaften nachweisbar auswirken, wird beim Experimentieren zumeist apparativ in einen zu erforschenden Ablauf kontrolliert eingegriffen um Ursache-Wirkungsbeziehungen generieren zu können. Beim Vergleichen werden stattdessen vorhandene Gemeinsamkeiten und Unterschiede kriteriengeleitet erarbeitet und dokumentiert. Bei der Durchführung müssen mögliche Störgrößen identifiziert und kontrolliert werden. Die Auswertung der Daten umfasst die Beurteilung der Datenqualität, ihre Analyse und Interpretation. Die Ergebnisse werden mit den Hypothesen verglichen und gedeutet. Beobachtungsdaten lassen Rückschlüsse auf korrelative Zusammenhänge zu, die Form- und Funktionsbeziehungen (z. B. Blüte - Art der Bestäubung), zeitliche Abfolgen (z. B. Sukzession, Phänologie, Ontogenese) oder räumliche Beziehungen

(z. B. Vegetation - Umweltfaktoren) betreffen. Kausalaussagen sind bei der Interpretation von Beobachtungsdaten unzulässig. Die Ergebnisse eines Vergleichs in Form von Gruppen oder hierarchischen Systemen ermöglichen phylogenetisch oder ökologisch bedingte Ähnlichkeiten aufzudecken (Homologien, Angepasstheiten). Beim Experimentieren können Ursachen eines Phänomens ermittelt werden, welche physiologischen, ökologischen, genetischen Ursprungs sein können.

4.2 Ausdifferenzierung von Kompetenzniveaus

Auf Basis der wissenschaftsmethodischen Binnenstruktur des vorliegenden Modells (Tab. 2) wurden für jede methodenspezifische Teilkompetenz (z. B. Fragestellung zum Experimentieren, Untersuchungsdesign zum Beobachten) strukturbezogene Kompetenzniveaus beschrieben. Dafür wurde zunächst das von Kauertz und Fischer (2006) beschriebene Aufgabenmerkmal *Komplexität* (z. B. 1 Fakt, 2 Fakten) zur Stufung von Aufgabenschwierigkeiten adaptiert (Tab. 3). Es wird als allgemeines Strukturmerkmal herangezogen, um Umfang und Vernetzungsgrad zu bearbeitender Inhalte zu operationalisieren. Mit diesem allgemeinen Strukturmerkmal ist es möglich, Testaufgaben für verschiedene Niveaus zu konstruieren, die im Idealfall modellkonforme Schwierigkeitsgrade aufweisen (Kauertz, 2008). Zur inhaltlichen, methodenspezifischen Beschreibung wurden die von Mayer et al. (2008) für das Experimentieren identifizierten Kompe-

Tab. 3: Kompetenzniveaus (I bis V) biologischer Erkenntnismethoden am Beispiel *Untersuchungsdesign*

	PLANUNG EINER UNTERSUCHUNG (Mayer et al., 2008)	UNTERSUCHUNGSDESIGN (Wellnitz, 2012)		
		Beobachten	Vergleichen	Experimentieren
Niveau V übergeordnetes Konzept*	Niveau V: Selbstständiges Lösen von offenen Problemen: Untersuchungsmethoden, Genauigkeit, Fehler abwägen	Eigenschaften der Beobachtung identifizieren z. B. keine Manipulation von Merkmalen, Selektion	Eigenschaften des Vergleichs identifizieren z. B. Trennschärfe, Kriterienstetigkeit, Auswahl mehrerer Vergleichskriterien	Eigenschaften des Experiments identifizieren z. B. Manipulation von Variablen, Variablenkontrolle
	Niveau IV: Elaborierte Untersuchung allgemeiner Zusammenhänge: Stichprobe, Messwiederholung und Dauer berücksichtigen			
Niveau IV 2 Zusammenhänge*	Niveau III: Kontrollierte Untersuchung auf Basis von Konzeptverständnis: Kontrollvariablen berücksichtigen	Ein System mit seinen Merkmalen und Veränderungen unter Berücksichtigung der Merkmalsausprägung in Beziehung setzen	Mehrere übergeordnete Vergleichskriterien unter Berücksichtigung ihrer Ausprägungen mit den Gruppen eines Ordnungssystems in Beziehung setzen	Eine abhängige und eine unabhängige Variable unter Berücksichtigung von Kontroll- u. Experimentalansatz bzw. Kontrollvariablen in Beziehung setzen
Niveau III 1 Zusammenhang*	Niveau II: Untersuchung von Zusammenhängen: veränderte und zu messende Variable in Beziehung setzen	Ein System mit seinen Merkmalen und Veränderungen in Beziehung setzen	Ein übergeordnetes Vergleichskriterium mit den Gruppen eines Ordnungssystems in Beziehung setzen	Eine abhängige und eine unabhängige Variable in Beziehung setzen
Niveau II 2 Fakten*		Identifizierung von zwei Merkmalen oder Veränderungen eines biologischen Systems	Identifizierung von zwei Unterschieden oder Gemeinsamkeiten von mindestens zwei biologischen Systemen	Identifizierung von zwei Variablen (abhängige, unabhängige oder Kontrollvariable) eines biologischen Phänomens
Niveau I 1 Fakt*	Niveau I: Untersuchung eines Faktors: eine Variable identifizieren	Identifizierung eines Merkmals, einer Struktur oder einer Veränderung eines biologischen Systems	Identifizierung eines Unterschiedes oder einer Gemeinsamkeit von mindestens zwei biologischen Systemen	Identifizierung einer abhängigen, einer unabhängigen oder einer Kontrollvariable eines biologischen Phänomens

* nach Kauertz & Fischer (2004)

tenzniveaus aufgegriffen und unter Berücksichtigung der Lernvoraussetzungen (siehe Kapitel 3) auf die Erkenntnismethoden *Beobachten* und *Vergleichen* übertragen. Die Synthese aus *Komplexität* (Kauertz, 2008) und *qualitativer Abstufung im Sinne eines problemlösenden Erkenntnisprozesses* (Mayer et al., 2008) erlaubt es, die Komplexität des zunehmenden wissenschaftsmethodischen Anspruchs für jedes Niveau, so zunächst die Annahme, abzubilden. Beispielhaft wird diese Vorgehensweise für die Teilkompetenz *Untersuchungsdesign* vorgestellt.

In Anlehnung an Kauertz und Fischer (2004) wird von den Schülerinnen und Schülern für das Erreichen des Niveaus I lediglich die Berücksichtigung eines Faktes erwartet (Tab. 3). Bei der Planung eines experimentellen Designs entspricht *1 Fakt* bei Mayer et al. (2008) der *Identifizierung einer Variablen*. Dementsprechend wurden die Niveaus in dieser Studie für das Experimentieren analog beschrieben (Niveau I: Identifizierung einer abhängigen, einer unabhängigen oder einer Kontrollvariable) und auf das Beobachten (Identifizierung eines Merkmals, einer Struktur oder einer Veränderung) und Vergleichen (Identifizierung eines Unterschiedes oder einer Gemeinsamkeit) theoriegeleitet übertragen (Tab. 3). *Zwei Zusammenhänge* zu verknüpfen (Kauertz, 2008) bzw. neben unabhängiger und abhängiger Variable auch *Kontrollvariablen* bei einem Experiment zu *berücksichtigen*, stellt eine größere Herausforderung für Lernende dar (Mayer et al., 2008). Ähnliches gilt für das Beobachten und Vergleichen, wo es Lernenden schwer fällt, Merkmale biolo-

gischer Systeme unter Berücksichtigung von Störgrößen in Beziehung zu setzen (Oğuz & Yurumezoğlu, 2007) bzw. mehrere Systeme kriterienstet und trennscharf zu sortieren (Hammann, 2002). Diese Vorgehensweisen entsprechen dem postulierten Niveau IV (Tab. 3).

Im Unterschied zu Schulleistungsstudien wie PISA oder TIMSS, bei denen üblicherweise erst im nachhinein die Aufgabenschwierigkeiten zur Beschreibung von Kompetenzniveaus herangezogen werden (Baumert, Bos & Lehmann, 2000; Prenzel, Häußler, Rost & Senkbeil, 2002), können somit a priori, also vor Datenerhebung, Hypothesen über wissenschaftsmethodische Anforderungen bei der Aufgabenentwicklung methodenspezifisch und prozessbezogen berücksichtigt werden (Wellnitz et al., 2012). Auf Basis der so abgeleiteten Binnenstruktur ist es möglich, die erforderlichen Teilkompetenzen für jede Erkenntnismethode konkret inhaltlich zu graduieren und in empirisch zu erfassende Konstrukte in Form von Testaufgaben zu überführen.

5 Zielsetzung

In der empirischen Untersuchung wurde der Frage nachgegangen, inwieweit sich die wissenschaftsmethodischen Konstrukte (Erkenntnisschritte und Erkenntnismethoden) und deren Graduierungen als kognitive Kompetenzkonstrukte in ihrer Struktur bestätigen lassen. Zur Prüfung der theoretisch hergeleiteten qualitativen Kompetenzanforderungen (Kompetenzniveaus) ist folgende Hypothese leitend:

H₁: Die mittleren Aufgabenschwierigkeiten steigen mit den zugrunde liegenden Kompetenzniveaus an.

Zur Prüfung der theoretisch hergeleiteten Teilkompetenzen (Kompetenzstruktur) werden drei konkurrierende Modelle miteinander verglichen, aus denen sich entsprechend drei Alternativhypothesen ergeben. Bei dem Obermodell *Naturwissenschaftliche Untersuchungen* (1 Dimension) werden die zu erfassenden wissenschaftsmethodischen Kompetenzen auf einem Kontinuum beschrieben. Es wird angenommen, dass alle Aufgaben auf einer Dimension laden:

H_{2,0}: Die Kompetenz *Naturwissenschaftliche Untersuchungen* lässt sich durch eine Dimension abbilden. Mit mehrdimensionalen Modellen wird keine bessere Passung erreicht.

Im Gegensatz dazu wird beim Modell *Erkenntnisschritte* (4 Dimensionen) vermutet, dass die prozessbezogenen Aspekte *Fragestellung*, *Hypothese*, *Untersuchungsdesign* und *Datenauswertung* eigenständige Dimensionen bilden:

H_{2,1}: Die Kompetenz *Naturwissenschaftliche Untersuchungen* lässt sich durch die vier Teilkonstrukte *Fragestellung*, *Hypothese*, *Untersuchungsdesign* und *Datenauswertung* abbilden und weist eine bessere Passung als das ein- oder das dreidimensionale Modell auf.

Beim Modell *Erkenntnismethoden* (3 Dimensionen) besteht die Annahme, dass

die drei Dimensionen *Beobachten*, *Vergleichen* und *Experimentieren* die empirischen Daten adäquat abbilden.

H_{2,2}: Die Kompetenz *Naturwissenschaftliche Untersuchungen* lässt sich durch die Teilkonstrukte *Beobachten*, *Vergleichen* und *Experimentieren* abbilden und weist eine bessere Passung als das ein- oder das vierdimensionale Modell auf.

6 Methodisches Vorgehen

6.1 Messinstrument

Die postulierte Kompetenzstruktur und deren Niveaus wurde im Rahmen des ES-NaS-Projekts (Wellnitz et al., 2012) mit Hilfe eines aufgabenbasierten Testinstruments (Papier-und-Bleistift-Test) überprüft. Von bundesweit abgeordneten Lehrkräften wurden 132 modellkonforme Aufgaben im geschlossenen, halboffenen und offenen Antwortformat konstruiert (Walpuski, Kampa, Kauertz & Wellnitz, 2008), die sich auf jeder graduierten Facette aus gekreuzten prozessbezogenen (*Fragestellung*, *Hypothese*, *Untersuchungsdesign* und *Datenauswertung*) und methodenspezifischen Teilkompetenzen (*Beobachten*, *Vergleichen*, *Experimentieren*) verorten lassen (Abb. 1, 2 und 3).

Beispielsweise müssen die Schülerinnen und Schüler beim naturwissenschaftlichen Beobachten im Aufgabenbeispiel „Wespennest“ (Abb. 1) die Notwendigkeit einer Blindprobe zur Gewährleistung der Funktionsfähigkeit des Nachweismittels

WESPENNEST

Philipp und Florian haben ein verlassenes Wespennest (Abb. 1) gefunden.

„Fühlt sich an wie Papier“, meint Philipp, als er das Nest vorsichtig aufhebt. Philipp und Florian lesen im Lexikon nach und finden heraus, dass Papier hauptsächlich aus Cellulose besteht. Von ihrer Lehrerin erfahren sie, dass es für Cellulose ein Nachweismittel gibt.

Sie tropfen das Nachweismittel auf ein Stück Papier und auf ein Stück des Wespennestes.

Erkläre, warum Philipp und Florian neben dem Wespennest auch ein Stück Papier mit dem Nachweismittel testen müssen.



© Liane Figge

Abb. 1: Wespennest

Abb. 1: Aufgabenbeispiel zum Beobachten (Niveau III): Teilkompetenz *Untersuchungsdesign* (IQB, 2013, S. 15 f).

HAIE UND ROCHEN

Obwohl Seekühe und Elefanten auf den ersten Blick recht unterschiedlich aussehen, sind beide miteinander verwandt. Dies geht auf Ähnlichkeiten in den Körpermerkmalen und Ähnlichkeiten auf molekularer Ebene zurück. Wenn sich Organismen in vielen Merkmalen gleichen, kann dies aber auch ein Zeichen von Anpasstheit an ähnliche Umweltbedingungen sein.

Zwei Wissenschaftler haben Haie und Rochen miteinander verglichen und festgestellt, dass beide schuppenartige Zähnnchen auf ihrer Haut (Placoidschuppen) besitzen. Sie deuten dies als Hinweis auf Verwandtschaft.

Erkläre, warum die Deutung der beiden Wissenschaftler nicht ausreichend ist.

Abb. 2: Modellkonformes Aufgabenbeispiel zum Vergleichen (Niveau V): Teilkompetenz *Datenauswertung* (Wellnitz, 2012, S. 67).

FLEISSIGES LIESCHEN

Max will die Wachstumsbedingungen von Pflanzen untersuchen. Er verwendet als Versuchspflanze das „Fleißige Lieschen“, eine pflegeleichte und widerstandsfähige Zimmerpflanze.

Er führt folgendes Experiment durch:

	Pflanze 1	Pflanze 2
Temperatur	25 °C	25 °C
Ort	am hellen Fenster	im dunklen Schrank
Gießen	1 mal täglich	1 mal täglich
Düngen	2 mal wöchentlich	2 mal wöchentlich

Nenne die Hypothese (Vermutung), die dem Experiment von Max zugrunde liegt.

Abb. 3: Aufgabenbeispiel zum Experimentieren (Niveau III): Teilkompetenz *Hypothese* (Wellnitz et al., 2012, S. 62)

darlegen können. Ohne die gleichzeitige Prüfung auf Papier wäre nicht ersichtlich, ob mit dem Reagenz Cellulose nachgewiesen werden kann.

6.2 Design und Stichprobe

Um die Testzeit für die Schülerinnen und Schüler möglichst gering zu halten, aber dennoch möglichst viele verschiedene Kompetenzkonstrukte untersuchen zu können, wurden die Testaufgaben, wie es in großen Schulleistungsstudien häufig der Fall ist, über ein sogenanntes Multi-Matrix-Design auf mehrere Testhefte verteilt (Wellnitz et al., 2012). Das bedeutet, dass die Schülerinnen und Schüler nicht alle Testaufgaben bearbeiten mussten und ihnen nur eine begrenzte Auswahl aus dem Aufgabenpool vorgelegt wurde.

Die Erhebung wurde vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) zur Erprobung (Pilotierung) der Bildungsstandardsaufgaben koordiniert und vom IEA Data Processing and Research Center (DPC) in Hamburg organisiert und durchgeführt. Die Erhebungen führten schulexterne Testleiterinnen und Testleiter durch, die durch das DPC geschult worden sind. An der Pilotierung haben insgesamt 6854 Schülerinnen und Schüler der 10. Jahrgangsstufe aus 160 Real- und Gesamtschulen sowie Gymnasien aus acht Bundesländern teilgenommen, denen neben Aufgaben aus der Biologie auch Items aus der Chemie und der Physik vorlagen (Wellnitz et al., 2012). Für die Beantwortung eigener Forschungsfragen wurden ausschließlich die Probanden aus-

gewählt ($N=986$), von denen vollständige Hintergrundmerkmale (z. B. Schulnoten, Schulform) vorhanden waren und denen mindestens ein Item aus der Biologie zu jeder Teilkompetenz (z. B. Beobachten, Hypothese) zur Bearbeitung vorlag.

6.3 Testtheoretischer Hintergrund

Die aus dem Multi-Matrix-Design resultierende Datenmatrix weist designbedingt eine große Anzahl fehlender Werte auf. Die Testhefte sind jedoch über sogenannte Ankeritems derart verlinkt, dass eine zuverlässige Schätzung der Aufgabenschwierigkeiten (Itemparameter) und Personenfähigkeiten (Personenparameter) mit Methoden der probabilistischen Testtheorie (Item-Response-Theorie), ermöglicht wird (Rauch & Hartig, 2008). Itemschwierigkeiten und Personenfähigkeiten werden auf einer gemeinsamen Logit-Skala abgebildet und können direkt aufeinander bezogen interpretiert werden. Die Schätzung der Itemschwierigkeiten und Personenfähigkeiten wurde mit der Software ConQuest von ACER durchgeführt (Wu, Adams, Wilson & Haldane, 2007).

6.4 Prüfung der Hypothesen

Zunächst wurden die Items dahingehend analysiert, ob deren Kennwerte denen in der probabilistischen Testtheorie üblichen Qualitätskriterien genügen. Die Schätzung der Itemparameter findet fallzentriert statt, indem der Mittelwert der Personen-

fähigkeiten auf Null gesetzt wird (Wu et al., 2007). Aus der Analyse resultieren sogenannte Item-Fits (wMNSQ, T-Wert), die Hinweise auf die Passung einzelner Items zu den jeweiligen Modellen geben.

Um zu prüfen, ob sich die gruppierten Items der Niveaus I bis V erwartungsgemäß in ihren Schwierigkeiten unterscheiden (H_1) wurden einfaktorielle Varianzanalysen (ANOVA) und Kontrastberechnungen durchgeführt. Da die Unterschiedshypothesen a priori formuliert und sowohl theoretisch begründet als auch empirisch fundiert sind, muss nach Bortz (1999) keine Alpha-Fehler-Korrektur vorgenommen werden. Zur Prüfung, ob die zugrundeliegenden Niveaumerkmale geeignet sind, die Schwierigkeit der Items von einem zum nächsten Niveau zu beeinflussen, wurden lineare Trends analysiert. Um den Einfluss der Niveaumerkmale auf die Itemschwierigkeiten zu quantifizieren, wurden multiple lineare Regressionsanalysen berechnet. Hierzu wurden die Niveaumerkmale, da es sich um ordinale Variablen handelt, als Dummy-Variablen kodiert (Hartig, 2007). Zudem wurden die Modelle auf Multikollinearität und Autokorrelation der Residuen überprüft.

Die Prüfung der postulierten Modelle (H_2) erfolgt mit dem Ziel, das Modell zu ermitteln, welches die beste Übereinstimmung mit den empirischen Daten aufweist. Abweichungen (Devianzen) zwischen postuliertem Modell und beobachteten Daten werden über den Vergleich der alternativen Modelle, d.h. über die Differenz der Devianzen betrachtet. Darüber hinaus findet die Modellkomplexität (Parameterzahl), die diese Passung erzielt, als

Freiheitsgrade (*df*: Differenz der Parameterzahl) Eingang in die Modellvergleiche. Die Modellprüfungen werden auf Basis der geschätzten Personenparameter durchgeführt (Wu et al., 2007). Als Entscheidungshilfe werden informationstheoretische Maße, in diesem Fall das Bayes Information Criterion (BIC), herangezogen (Bühner, 2006). Das BIC eignet sich im Vergleich zum Consistent Akaike Information Criterion (CAIC) besonders für lange Tests mit kleinen Häufigkeiten der beobachteten Antwortmuster und zieht im Gegensatz zum Akaike Information Criterion (AIC) die Stichprobengröße zur Berechnung der Indizes heran (Rost, 2004). Da sich die informationstheoretischen Maße nicht absolut, sondern nur durch den Vergleich konkurrierender Modelle interpretieren lassen, werden die Alternativmodelle hinsichtlich ihrer Passung auf die Daten zusätzlich mit dem Likelihoodquotiententest auf Signifikanz geprüft. Für die Durchführung des Likelihoodquotiententest werden von Rost (2004) folgende Bedingungen formuliert, die als erfüllt betrachtet werden: (a) Alle Modelle basieren auf der gleichen Datenmatrix, (b) das eindimensionale Modell *Naturwissenschaftliche Untersuchungen* ist ein echtes Obermodell und durch Restriktion der Parameter in die beiden mehrdimensionalen Alternativmodelle *Erkenntnissschritte* und *Erkenntnismethoden* überführbar und (c) die restriktiveren Modelle gehen nicht durch Null-Setzen einzelner Parameter aus dem Obermodell hervor.

7 Ergebnisse

Von den ursprünglich 132 Aufgaben entsprechen 116 den in der probabilistischen Testtheorie üblichen Qualitätskriterien ($0.80 \leq \text{MNSQ} \leq 1.20$; $T\text{-Werte} \leq 1.96$) (Adams, 2002; Rost, 2004). Die Aufgabenschwierigkeiten der Items liegen innerhalb des angestrebten Wertebereichs von -3 Logits bis +3 Logits (Rauch & Hartig, 2008). Die mittlere Aufgabenschwierigkeit liegt bei -0.46 ($SD = 1.21$). Damit sind die nach der Itemselektion ausgewerteten Testaufgaben für die Schülerinnen und Schüler der 10. Jahrgangsstufe tendenziell etwas zu leicht. Der Test weist insgesamt eine EAP/PV-Reliabilität von 0.59 auf.

7.1 Prüfung der Kompetenzniveaus

Die mit einer einfaktoriellen Varianzanalyse durchgeführte Prüfung der Nullhypothese, dass die Items aller fünf Kompetenzniveaus gleich schwer sind, wird mit einem starken signifikanten Effekt von $\eta_p^2 = .28$ verworfen ($F(4, 111) = 10.55$, $p < .001$). Dabei unterliegt die Unterschiedlichkeit der Mittelwerte einem signifikanten linearen Trend ($F(1, 111) < 38.25$, $p < .001$, $\omega p^2 < .25$). Abbildung 4 zeigt jedoch, dass der Anstieg der Itemschwierigkeiten nicht durchgängig von einem Niveau zum nächsten verläuft.

Werden die Mittelwertunterschiede zwischen benachbarten Niveaus des

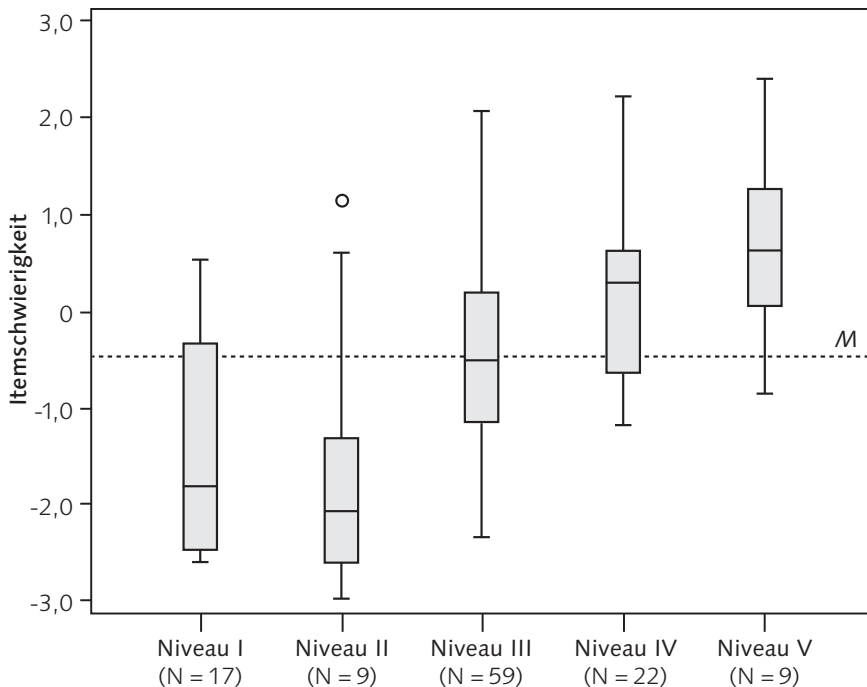


Abb. 4: Boxplots der Itemschwierigkeiten. Jede Box zeigt die beiden Quartile und den Median an. Der kleine Kreis repräsentiert einen Ausreißer (M: Mittelwert).

Tab. 4: Kontraste der Itemparameter-Mittelwerte

	<i>M</i>	<i>SD</i>	Niveau I	Niveau II	Niveau III	Niveau IV
Niveau I	-1.42	1.09				
Niveau II	-1.52	1.46	$\Delta M_{II,I} = -0.10$			
Niveau III	-0.44	1.03	$\Delta M_{III,I} = 0.99^{**}$	$\Delta M_{III,II} = 1.08^{**}$		
Niveau IV	0.16	0.88	$\Delta M_{IV,I} = 1.58^{***}$	$\Delta M_{IV,II} = 1.68^{***}$	$\Delta M_{IV,III} = 0.60^*$	
Niveau V	0.72	1.05	$\Delta M_{V,I} = 2.13^{***}$	$\Delta M_{V,II} = 2.24^{***}$	$\Delta M_{V,III} = 1.16^{**}$	$\Delta M_{V,IV} = 0.55$

p* < .05. ** *p* < .01. * *p* < .001.

Tab. 5: Regressionsgewichte zur Vorhersage der Itemschwierigkeit (*B*: unstandardisiertes Regressionsgewicht, *SE_B*: Standardfehler des unstandardisierten Regressionsgewichtes, β : standardisiertes Regressionsgewicht)

NIVEAUS	BEOBACHTEN			VERGLEICHEN			EXPERIMENTIEREN		
	<i>B</i>	<i>SE_B</i>	β	<i>B</i>	<i>SE_B</i>	β	<i>B</i>	<i>SE_B</i>	β
Regressionskonstante <i>B₀</i>	-0.47	0.47		-1.32	0.56		-1.82	0.25	
Niveau III	0.13	0.52	.06	0.80	0.70	.33	1.29	0.33	.48***
Niveau IV	0.81	0.65	.27	1.64	1.38	.33	1.88	0.36	.64***
Niveau V	1.11	0.90	.23	1.40	1.38	.28	2.68	0.48	.63***
Varianzaufklärung	$R^2 = .09$, $R^2_{\text{kor}} = .01$ (<i>p</i> = .330)			$R^2 = .17$, $R^2_{\text{kor}} = -.04$ (<i>p</i> = .513)			$R^2 = .44$, $R^2_{\text{kor}} = .41$ (<i>p</i> < .001)		

*** *p* < .001.

Obermodells *Naturwissenschaftliche Untersuchungen* mit Kontrastanalysen auf ihre Signifikanz getestet, zeigt sich, dass sich die Niveaumerkmale weitestgehend in den Aufgabenschwierigkeiten widerspiegeln (Tab. 4). Die Items der fünf Niveaus differieren zwischen Niveau II und Niveau III sowie zwischen Niveau III und Niveau IV signifikant in ihren Schwierigkeiten. Die mittlere Aufgabenschwierigkeit nimmt von *Niveau II* bis *Niveau V* modellkonform zu. Die Items der Niveaus I und II sind hingegen annähernd gleich schwer.

Für weiterführende Regressionsanalysen wurden, um den Einfluss der Niveaumerkmale auf die Itemschwierigkeit getrennt für die Erkenntnismethoden *Beobachten*, *Vergleichen* und *Experimentieren* untersuchen zu können, die Niveaus I und II als zusammengefasste Referenzgruppe definiert (Regressionskonstante *B₀*). Die Regressionskonstante *B₀* entspricht bei allen Erkenntnismethoden mit -0.47 (*Beobachten*), -1.32 (*Vergleichen*) und -1.82 (*Experimentieren*) der erwarteten Schwierigkeit einfachster Aufgabe der Niveaus I und II (Tab. 5).

Die Werte der unstandardisierten Regressionsgewichte zeigen, dass die Items zum Beobachten und Experimentieren modellkonform von einem Niveau zum nächsten schwerer werden. Signifikante Beiträge zur Varianzaufklärung liegen für die Erkenntnismethode *Experimentieren* vor. Hier besitzen die Aufgabenmerkmale des *Niveaus III* ($\beta = .48$), des *Niveaus IV* ($\beta = .64$) und des *Niveaus V* ($\beta = .63$) hohe standardisierte Regressionsgewichte, die einen erwartungsgemäß starken Einfluss auf die Schwierigkeiten der Items repräsentieren (Tab. 5). Insgesamt klären die niveaubezogenen Prädiktoren der Erkenntnismethode *Experimentieren* 44 % an Gesamtvarianz der Itemschwierigkeiten auf. Die nicht-signifikanten Determinationskoeffizienten der Erkenntnismethoden *Beobachten* und *Vergleichen* verweisen darauf, dass der Einfluss der Niveaumerkmale auf die Itemschwierigkeiten nicht zuverlässig vorhergesagt werden kann. Bei beiden Erkenntnismethoden sind jedoch Effekte der Niveaus auf die Itemschwierigkeit erkennbar, die sich in höheren Regressionsgewichten niederschlagen.

7.2 Prüfung der Kompetenzstruktur

Das Devianzmaß des restringierten dreidimensionalen Modells *Erkenntnismethoden* (3 Dimensionen) fällt niedrigerer aus als die Werte des ein- und des vierdimensionalen Modells (Tab. 6). Da die Devianz ein Maß für die Abweichung eines Modells von den empirischen Daten darstellt, ist damit die Abweichung des dreidimensionalen Modells am geringsten. Das Modell

Erkenntnisschritte (4 Dimensionen) weist wiederum ein niedrigeres Devianzmaß als das Modell *Naturwissenschaftliche Untersuchungen* (1 Dimension) auf (Tab. 6).

Beim BIC werden die Devianzmaße durch die Anzahl der Modellparameter unter Berücksichtigung der Stichprobengröße relativiert (Rost, 2004). Nach den informationstheoretischen Maßen des BIC entspricht das eindimensionale Modell *Naturwissenschaftliche Untersuchungen* am besten den Daten ($BIC_{1D} < BIC_{3D} < BIC_{4D}$). Dies wird durch das informationstheoretische Maß von 18194.52 verdeutlicht, welches im Vergleich zu den beiden Alternativmodellen *Erkenntnismethoden* ($\Delta = 25.29$) und *Erkenntnisschritte* ($\Delta = 56.34$) niedriger ausfällt (Tab. 6). Der Vergleich zwischen den Modellen *Erkenntnismethoden* und *Erkenntnisschritte* fällt zugunsten des dreidimensionalen Modells aus ($\Delta = 31.05$). Zur genaueren Analyse erfolgt deshalb eine zusätzliche Prüfung der Modellannahmen mit dem Likelihoodquotiententest (Tab. 7). Die Prüfgröße des Likelihoodquotienten lässt sich aus der Differenz der Devianzmaße zweier Modelle berechnen und in eine χ^2 -verteilte Prüfstatistik umwandeln (Rost, 2004). Aus Tabellen für Chi-Quadrat-Verteilungen kann unter Berücksichtigung der Freiheitsgrade (df) der kritische Wert für die Prüfgröße (χ^2_{krit}) des gewünschten Signifikanzniveaus entnommen werden.

Die Analyse der Modellgeltungstests ergibt für die Gegenüberstellung aller drei konkurrierenden Modelle, dass die Differenzen zwischen den jeweiligen Modellen statistisch nicht signifikant sind. Zwar weist das Modell *Erkenntnismetho-*

Tab. 6: Modellprüfung mit dem BIC

	DIMENSIONEN	PARAMETERZAHL	DEVIANZMASS	BIC
Naturwissenschaftliche Untersuchungen	1 D	117	17388.39	18194.52
Erkenntnismethoden	3 D	122	17379.23	18219.81
Erkenntnisschritte	4 D	126	17382.72	18250.86

Tab. 7: Modellprüfung mit dem Likelihoodquotiententest (D: Dimensionen)

	df	χ^2_{emp}	χ^2_{krit}	p
1 D/3 D	5	9.16	11.07	> .05
1 D/4 D	9	5.67	16.92	> .05
3 D/4 D	4	3.49	9.49	> .05

den das niedrigste Devianzmaß auf, die Unterschiede zu den beiden Alternativmodellen sind jedoch nicht bedeutend. Bei der vorliegenden Anzahl von fünf Freiheitsgraden (*df*, Differenz der Parameterzahl) zwischen den Modellen *Naturwissenschaftliche Untersuchungen* und *Erkenntnismethoden* wäre beispielsweise mindestens eine Differenz der Devianzmaße von 11.07 notwendig, um einen Unterschied zwischen beiden Modellen belegen zu können. Die Differenz beträgt 9.16. Der Unterschied ist dementsprechend nicht signifikant ($p > .05$).

8 Interpretation

8.1 Eignung des Testinstruments

Dieser Studie lagen schriftliche Testaufgaben zugrunde. Naturwissenschaftliche Erkenntnisgewinnung ist in der Regel aber

an die praktische Durchführung von Untersuchungen zur Datengewinnung gebunden (Meier & Mayer, 2012). Inwieweit mit schriftlicher Leistungsmessung, wie es in dieser Studie der Fall ist, auch praktische Fähigkeiten erfasst werden können, konnte bisher nicht hinreichend geklärt werden (Emden, 2011; Schreiber, Theyßen & Schecker, 2009; Shavelson et al., 1991). Zwar wurde in den Aufgaben dieser Studie keine praktische Umsetzung gefordert, allerdings sollte diese von den Schülerinnen und Schülern beschrieben und bezüglich der Angemessenheit beurteilt werden. Alle Aufgaben haben das im ESNaS-Projekt übliche Verfahren zur Qualitätssicherung durchlaufen (Pant, Stanat, Pöhlmann & Böhme, 2013). Die Standardisierung der verschiedenen Schritte von der Itemkonstruktion bis hin zu deren Testung wird im Allgemeinen als ausreichend für eine hinreichende Testobjektivität angesehen (Rost, 2004). Die Passung der Aufgaben-

daten zu den geschätzten Modellparametern ist bei allen drei Modellannahmen (*Naturwissenschaftliche Untersuchungen, Erkenntnismethoden, Erkenntnisschritte*) gegeben. Dies spricht für eine hohe interne Konsistenz, d.h. dass alle zugrundeliegenden Items im Wesentlichen ein gemeinsames Konstrukt messen. Die Items sind Rasch-homogen.

Die EAP/PV-Reliabilität des Obermodells *Naturwissenschaftliche Untersuchungen* ist mit 0.59 vergleichsweise gering. Eine Ursache hierfür könnte darin liegen, dass den Schülerinnen und Schülern im Rahmen der ESNaS-Pilotierungsstudie, neben Aufgaben aus der Chemie und der Physik, im Mittel 16 biologische Items aus dem Bereich *Naturwissenschaftliche Untersuchungen* zur Bearbeitung vorgelegt wurden ($SD = 6.51$; $Min = 5$, $Max = 31$). Bei Bearbeitung geringer Itemzahlen muss mit individuellen Messungenauigkeiten gerechnet werden, die wiederum die Reliabilitäten der Skalen beeinflussen (Adams, 2005). Ein weiterer Grund für die verhältnismäßig geringen Reliabilitäten könnte im Zusammenhang mit dem hohen Anteil verwendeter offener Antwortformate stehen (54%), die zugunsten der selbstständigen Generierung von Lösungswegen verwendet wurden. Mit zunehmendem Anteil offener Antwortformate gehen die Reliabilitäten von Gesamtskalen in der Regel zurück (Stecher & Klein, 1997).

8.2 Kompetenzniveaus

Die für das Obermodell *Naturwissenschaftliche Untersuchungen* durchge-

führten Schwierigkeitsanalysen zeigen, dass die a priori formulierten Niveaumerkmale grundsätzlich für die Itemkonstruktion geeignet sind, um die Komplexität des wissenschaftsmethodischen Anspruchs abzubilden. Dieses Ergebnis knüpft an naturwissenschaftliche Studien an, die ebenfalls die Komplexität als schwierigkeitsbestimmendes Aufgabenmerkmal postuliert haben. Der Anstieg der mittleren Itemschwierigkeiten von einem Niveau zum nächsten konnte bereits für Testaufgaben zum chemischen Fachwissen für alle fünf Niveaus (Walpuski et al., 2011), zur naturwissenschaftlichen Erkenntnisgewinnung in der Chemie für die Niveaus I, III und IV (Mannel, 2011), zum physikalischen Fachwissen für die Niveaus I, II, III und V (Kauertz, 2008) und zur Bewertungskompetenz in Biologie und Chemie für die Niveaus III, IV und V (Hostenbach, 2011) bestätigt werden. Während Kauertz (2008) und Ropohl (2010) auch den Anstieg der Itemschwierigkeiten von Niveau I zu Niveau II aufzeigen konnten, sind die Schwierigkeiten der zugehörigen Testaufgaben dieser Studie erwartungswidrig ausgefallen. Das liegt möglicherweise an der mangelnden Itemsignierung des Niveaus II und den vergleichsweise hohen Standardabweichungen. Werden die Analysen getrennt für die Testaufgaben zum Beobachten, Vergleichen und Experimentieren durchgeführt, ist ebenfalls, mit Ausnahme der Items zum Vergleichen, ein modellkonformer Anstieg der mittleren Itemschwierigkeiten belegbar. Der schwierigkeitserzeugende Einfluss zeigt sich insbesondere dann, wenn die verschiedenen Niveaus durch eine vergleichsweise hohe

Itemzahl, wie es beim Experimentieren der Fall ist, repräsentiert werden. Schülerinnen und Schüler der 10. Jahrgangsstufe können Testaufgaben zum Experimentieren signifikant leichter lösen, wenn lediglich ein oder zwei Variablen prozessbezogen bearbeitet werden müssen. Das richtige Lösen von Aufgaben, in denen ein kausaler Zusammenhang kognitiv verarbeitet werden soll, fällt den Probanden demgegenüber signifikant schwerer. Aufgaben, in denen zwei Zusammenhänge methodenspezifisch und prozessbezogen hergestellt werden müssen, sind nochmals signifikant schwerer für die Schülerinnen und Schüler zu lösen. Der schwierigkeits-erzeugende Einfluss wissenschaftsmethodischer Kriterien erlaubt es, gezielt die Anforderungen zum Experimentieren in Testaufgaben zu beeinflussen.

8.3 Kompetenzstruktur

Es zeigt sich anhand der informationstheoretischen Maße des BIC, dass aus psychometrischer Sicht das eindimensionale Modell *Naturwissenschaftliche Untersuchungen* gegenüber den beiden Alternativmodellen *Erkenntnismethoden* und *Erkenntnisschritte* zu bevorzugen wäre. Die Differenzwerte zwischen allen Modellen sind größer als zehn, was von Burnham und Anderson (2004) als ausreichender Schwellenwert angegeben wird. Die inferenzstatistische Absicherung über Likelihoodquotiententests zeigt jedoch, dass alle Modelle gleichermaßen geeignet sind, die empirischen Daten abzubilden. Eine anspruchsvolle Ausdifferenzierung

des eindimensionalen Kompetenzkonstruktes *Naturwissenschaftliche Untersuchungen* in drei (*Beobachten, Vergleichen, Experimentieren*) bzw. vier Konstrukte (*Fragestellung, Hypothese, Untersuchungsdesign, Datenauswertung*) führt zu keiner signifikant schlechteren Anpassungsgüte. Auch wenn das eindimensionale Kompetenzkonstrukt nach dem wissenschaftstheoretischen Prinzip der Sparsamkeit bzw. der Einfachheit bevorzugt werden müsste (Rost, 2004), wird mit der theoretisch begründeten Ausdifferenzierung eine höhere Inhaltsvalidität (bezogen auf die naturwissenschaftliche Erkenntnisgewinnung) und gleichzeitig eine höhere ökologische Validität (bezogen auf „Schulversuche“) sichergestellt. Die in dieser Studie vorgestellte Ausdifferenzierung eines komplexen Modells naturwissenschaftlicher Erkenntnisgewinnung, bezieht den aktuellen Forschungsstand mit ein und berücksichtigt den Geltungsbereich zugrunde gelegter Theorien.

9 Fazit

Erstmals konnte für das Unterrichtsfach Biologie ein Kompetenzstrukturmodell vorgestellt werden, welches über das Experimentieren hinaus (Hamann et al., 2007; Klos et al., 2008; Mayer et al., 2008), auch die Operationalisierung weiterer Erkenntnismethoden wie das Beobachten und Vergleichen ermöglicht. Die wissenschaftsmethodische Ausdifferenzierung der Erkenntnismethoden eignet sich, neben dem Assessment, als Strukturraster im Rahmen von Aus- und Fortbildungs-

maßnahmen oder direkt für den Transfer in den Biologieunterricht. So wäre es möglich, die wissenschaftsmethodische Struktur originärer Erkenntnismethoden der Biologie nachvollziehbar in angemessenen Lehr-Lernarrangements umzusetzen und ein Verständnis für die Vielfalt biologischer Denk- und Arbeitsweisen zu fördern. Darüber hinaus kann die Ausdifferenzierung naturwissenschaftlicher Erkenntnismethoden auch konkret für den Unterricht zur Konzeption von Lern- oder Testaufgaben herangezogen werden. Dadurch, dass die zu erfassenden Kompetenzkonstrukte als zweidimensionales Facettendesign abgebildet werden, ist es trotz des hohen Differenzierungsgrades möglich, mit einer vergleichsweise geringen Itemzahl verschiedene Teilkompetenzen abzubilden.

Literatur

- Abd-El-Khalick, F. & Lederman, N. G. (2000). Improving Science Teachers' Conceptions of Nature of Science: A Critical Review of the Literature. *International Journal of Science Education*, 22(7), 665–701.
- Adams, R. J. (2002). Scaling PISA Cognitive Data. In R. Adams & M. Wu (Hrsg.), *PISA 2000: Technical Report* (S. 99–108). Paris: OECD.
- Adams, R. J. (2005). Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, 31, 162–172.
- Ahtee, M., Suomela, L., Juuti, K., Lampiselkä, J. & Lavonen, J. (2009). Primary school student teachers' views about making observations. *Nordic Studies in Science Education*, 5(2), 128–141.
- Arnold, J., Kremer, K. & Mayer, J. (2012). Wissenschaftliches Denken beim Experimentieren – Kompetenzdiagnose in der Sekundarstufe II. In D. Krüger, A. Upmeyer zu Belzen, P. Schmiemann, A. Möller & D. Elster (Hrsg.), *Erkenntnisweg Biologiedidaktik 11* (S. 7–20). Kassel: Universitätsdruckerei.
- Arnold, J., Wellnitz, N. & Mayer, J. (2011). Beschreibung und Messung von Beobachtungskompetenz bei Schülerinnen und Schülern der Sekundarstufe I. In D. Krüger, A. Upmeyer zu Belzen & S. Nitz (Hrsg.), *Erkenntnisweg Biologiedidaktik 9* (S. 7–22). Neumünster: Universitätsdruckerei Kassel.
- Aydeniz, M., Baksa, K. & Skinner, J. (2011). Understanding The Impact of an Apprenticeship-Based Scientific Research Program on High School Students' Understanding of Scientific Inquiry. *Journal of Science Education and Technology*, 20, 403–421.
- Baumert, J., Bos, W. & Lehmann, R. (Hrsg.) (2000). *TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit*. Opladen: Leske + Budrich.
- Bell, R. L., Blair, L. M., Crawford, B. A. & Lederman, N. G. (2003). Just Do It? Impact of a Science Apprenticeship Program on High School Students' Understanding of the Nature of Science and Scientific Inquiry. *Journal of Research in Science Teaching*, 40(5), 487–509.
- Boerwinkel, D. J., Waarlo, A. J. & Boersma, K. (2009). A designer's view: The perspective of form and function. *Journal of Biological Education*, 44(1), 12–18.
- Bortz, J. (1999). *Statistik für Sozialwissenschaftler*. Berlin: Springer.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin: Springer.
- Braund, M. (1991). Children's ideas in classifying animals. *Journal of Biological Education*, 25(2), 103–110.

- Burnham, K. P. & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33, 261–304.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.
- Carrier, M. (2005). Multiplizität und Heterogenität: Zur begrifflichen Eigenständigkeit biologischer Funktionen. In M. Stöltzner & P. Weingartner (Hrsg.), *Formale Teleologie und Kausalität in der Physik. Zur philosophischen Relevanz des Prinzips der kleinsten Wirkung und seiner Geschichte* (S. 297–310). Paderborn: Mentis.
- Chinn, C. A. & Brewer, W. F. (1998). An Empirical Test of a Taxonomy of Responses to Anomalous Data in Science. *Journal of Research in Science Teaching*, 35(6), 623–654.
- Darian, S. (1997). The Language of Experiments in Introductory Science Texts. *Fachsprache* 19, 28–42.
- DfES & QCA [Department for Education and Skills & Qualification and Curriculum Authority] (2004). *Science – The National Curriculum for England*. London: HMSO.
- Doran, R. L., Helgeson, S. L. & Kumar, D. D. (1995). Assessment of Problem Solving in Science. In D. R. Lavoie (Hrsg.), *Toward a Cognitive-Science Perspective for Scientific Problem Solving* (S. 112–141). Manhattan, Kansas: NARST Monograph, Number Six.
- Duggan, S., Johnson, P. & Gott, R. (1996). A Critical Point in Investigative Work: Defining Variables. *Journal of Research in Science Teaching*, 33(5), 461–474.
- Eichberg, E. (1972). *Über das Vergleichen im Unterricht*. Hannover: Schroedel.
- Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*. Berlin: Logos.
- Emden, M. & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung. Zur Eignung einer Protokollmethode zur Bewertung von Experimentierprozessen. *Der mathematische und naturwissenschaftliche Unterricht*, 65(2), 68–75.
- Flannery, M. C. (2007). Observations on Biology. *The American Biology Teacher*, 69(9), 561–564.
- George, K. D. & Dietz, M. A. (1971). How Do Children Classify Objects? *Journal of Research in Science Teaching*, 8(3), 277–283.
- Germann, P. J., Aram, R. & Burke, G. (1996). Identifying Patterns and Relationships among the Responses of Seventh-Grade Students to the Science Process Skill of Designing Experiments. *Journal of Research in Science Teaching*, 33(1), 79–99.
- Graßhoff, G., Casties, R. & Nickelsen, K. (2000). *Zur Theorie des Experiments – Untersuchungen am Beispiel der Entdeckung des Harnstoffzyklus*. Bern: Studies in the History and Philosophy of Science.
- Greve, W. & Wentura, D. (1997). *Wissenschaftliche Beobachtung – Eine Einführung*. Weinheim: Beltz Psychologie Verlags Union.
- Gropengießer, H. (2013). Beobachten. In H. Gropengießer, U. Harms & U. Kattmann (Hrsg.), *Fachdidaktik Biologie* (S. 273–277). Hallbergmoos: Aulis.
- Grube, C. (2011). Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung. Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I. Verfügbar unter <https://kobra.bibliothek.uni-kassel.de/handle/urn:nbn:de:hebis:34-2011041537247>. [17.06.2011].
- Hammann, M. (2002). *Kriteriengeleitetes Vergleichen im Biologieunterricht*. Innsbruck: StudienVerlag.
- Hammann, M. (2004a). Kompetenzentwicklungsmodelle. *Der mathematische und naturwissenschaftliche Unterricht*, 57(4), 196–203.
- Hammann, M. (2004b). Tiere ordnen – Ein Methodentraining zum kriteriengeleiteten Vergleichen. In R. Duit, H. Gropengießer & L. Stäudel (Hrsg.), *Naturwissenschaftliches Arbeiten – Unterricht und Material 5 – 10* (S. 38–46). Seelze-Velber: Friedrich Verlag.

- Hammann, M. (2005). Kompetenzentwicklungsmodelle für das kriteriengeleitete Vergleichen. In H. Bayrhuber, S. Bögeholz, D. Graf, M. Hammann, U. Harms, C. Hößle et al. (Hrsg.), *Bildungsstandards Biologie* (S. 47–50). Bielefeld: Internationale Tagung der Sektion Biologiedidaktik im VDBiol, 27. Februar bis 4. März.
- Hammann, M. & Bayrhuber, H. (2003). How do Students use Criteria in Comparisons? In J. Lewis, A. Magro & L. Simonneaux (Hrsg.), *Biology Education for the Real World. Student – Teacher – Citizen* (S. 259–272). Toulouse-Auzeville: Proceedings of the IVth ERIDOB Conference, Ecole nationale de formation agronomique.
- Hammann, M., Phan, T. H. & Bayrhuber, H. (2007). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? *Zeitschrift für Erziehungswissenschaften, Sonderheft 8*, 33–49.
- Hammann, M., Phan, T. T. H., Ehmer, M. & Bayrhuber, H. (2006). Fehlerfrei Experimentieren. *Der mathematische und naturwissenschaftliche Unterricht*, 59(5), 292–299.
- Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing Pupils' Skills in Experimentation. *Journal of Biological Education*, 42(2), 66–72.
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 83–99). Weinheim: Beltz.
- Harwood, W. S. (2004). A New Model for Inquiry. Is the Scientific Method Dead? *Journal of College Science Teaching*, 33(7), 29–33.
- Hostenbach, J. (2011). *Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht*. Berlin: Logos.
- IQB [Institut zur Qualitätsentwicklung im Bildungswesen] (Hrsg.) (2013). Kompetenzstufenmodelle zu den Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Kompetenzbereiche „Fachwissen“ und „Erkenntnisgewinnung“. Verfügbar unter <http://www.iqb.hu-berlin.de/bista/ksm> [04.10.2013].
- Janich, P. & Weingarten, M. (1999). *Wissenschaftstheorie der Biologie*. München: Wilhelm Fink.
- Kattmann, U. & Schmitt, A. (1996). Elementares Ordnen: Wie Schüler Tiere klassifizieren. *Zeitschrift für Didaktik der Naturwissenschaften*, 2(2), 21–38.
- Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. Berlin: Logos.
- Kauertz, A. & Fischer, H. E. (2006). Assessing Students' Level of Knowledge and Analysing the Reasons for Learning Difficulties in Physics by Rasch Analysis. In X. Liu & W. J. Boone (Hrsg.), *Applications of Rasch Measurement in Science Education* (S. 212–246). Maple Grove, USA: JAM.
- Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, 16, 135–153.
- Killermann, W., Hiering, P. & Starosta, B. (2005). *Biologieunterricht heute. Eine moderne Fachdidaktik*. Donauwörth: Auer.
- Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge: MIT.
- Klahr, D., Fay, A. L. & Dunbar, K. (1993). Heuristics for Scientific Experimentation: A Developmental Study. *Cognitive Psychology*, 25(1), 111–146.
- Klix, F. (1996). Lernen und Denken. In J. Hoffmann & W. Kintsch (Hrsg.), *Lernen* (S. 529–580). Hogrefe: Verlag für Psychologie.

- Klos, S., Henke, C., Kieren, C., Walpuski, M. & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen – zwei verschiedene Kompetenzen. *Zeitschrift für Pädagogik*, 54(3), 304–321.
- KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland] (Hrsg.) (2005). *Beschlüsse der Kultusministerkonferenz – Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. München, Neuwed: Luchterhand.
- Köchy, K. (2006). Lebewesen im Labor. Das Experiment in der Biologie. *Philosophia naturalis*, 43(1), 75–100.
- Köhler, K. (2004). Welche fachgemäßen Arbeitsweisen werden im Biologieunterricht eingesetzt? In U. Spörhase-Eichmann & W. Ruppert (Hrsg.), *Biologiedidaktik. Praxishandbuch für die Sekundarstufe I und II* (S. 146–159). Berlin: Cornelsen Scriptor.
- Kohlhauf, L. (2012). *Spielend Biologisch Beobachten. Messung und Förderung biologischer Beobachtungskompetenz im Vorschulalter*. Berlin: Logos.
- Kohlhauf, L., Rutke, U. & Neuhaus, B. (2011). Entwicklung eines Kompetenzmodells zum biologischen Beobachten ab dem Vorschulalter. *Zeitschrift für Didaktik der Naturwissenschaften*, 17, 203–222.
- Krüger, D. & Burmester, A. (2005). Wie Schüler Pflanzen ordnen. *Zeitschrift für Didaktik der Naturwissenschaften*, 11, 85–102.
- Lederman, N. G. (2004). Syntax of Nature of Science within Inquiry and Science Instruction. In L. B. Flick & N. G. Lederman (Hrsg.), *Scientific Inquiry and Nature of Science. Implications for Teaching, Learning, and Teacher Education* (S. 301–317). Dordrecht: Kluwer.
- Mahner, M. & Bunge, M. (2000). *Philosophische Grundlagen der Biologie*. Berlin: Springer.
- Mannel, S. (2011). *Assessing scientific inquiry. Development and evaluation of a test for the low-performing stage*. Berlin: Logos.
- Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Handbuch der Theorien in der biologiedidaktischen Forschung – Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 178–186). Berlin: Springer.
- Mayer, J., Grube, C. & Möller, A. (2008). Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In U. Harms & A. Sandmann (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 3. Ausbildung und Professionalisierung von Lehrkräften* (S. 63–79). Innsbruck: StudienVerlag.
- Mayr, E. (1984). *Die Entwicklung der biologischen Gedankenwelt. Vielfalt, Evolution und Vererbung*. Berlin: Springer.
- Meier, M. & Mayer, J. (2012). Experimentierkompetenz praktisch erfassen – Entwicklung und Validierung eines anwendungsbezogenen Aufgabendesigns. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik, Band 5* (S. 81–98). Innsbruck: StudienVerlag.
- McComas, W. F. (1996). Ten Myths of Science: Re-examining what we think we know about the Nature of Science. *School Science & Mathematics*, 96(1), 10–15.
- Meier, M. & Mayer, J. (2012). Experimentierkompetenz praktisch erfassen – Entwicklung und Validierung eines anwendungsbezogenen Aufgabendesigns. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 5* (S. 81–98). Innsbruck: StudienVerlag.
- Meier, M. & Wellnitz, N. (2013). Beobachten, Vergleichen und Experimentieren mit Wasserflößen. Biologische Erkenntnismethoden praktisch anwenden. *Praxis der Naturwissenschaft – Biologie in der Schule*, 1(62), 4–10.
- Möller, A., Hartmann, S. & Mayer, J. (2010). *Differentiation and development of five levels in scientific inquiry skills: a longitudinal assessment of Biology students in grade 5 to 10*. Paper presented at the Annual Meeting of the National Association of Research in Science Teaching (NARST), Philadelphia, PA, United States.

- NRC [National Research Council] (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academy Press.
- Norris, S. P. (1984). Defining Observational Competence. *Science Education*, 68(2), 129–142.
- Oğuz, A. & Yurumezoğlu, K. (2007). *The Primacy of Observation in Inquiry-based Science Teaching*. Paper presented at the International Workshop Science Education in School. Bucharest, Romania, Oct, 11–14.
- Oğuz-Ünver, A. & Yurumezoğlu, K. (2009). A Teaching Strategy for Developing the Power of Observation in Science Education. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi*, 28, 105–119.
- Pant, H. A., Stanat, P., Pöhlmann, C. & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 13–21). Münster: Waxmann.
- Popper, K. (1994). *Logik der Forschung*. Tübingen: Mohr.
- Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, 30(1), 120–135.
- Rauch, D. & Hartig, J. (2008). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 240–250). Berlin: Springer.
- Ropohl, M. (2010). *Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion: Entwicklung und Analyse von Testaufgaben*. Berlin: Logos.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.
- Ryman, D. (1974). Children's Understanding of the Classification of Living Organisms. *Journal of Biological Education*, 8(3), 140–144.
- Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen?! *Physik und Didaktik in Schule und Hochschule*, 8(3), 92–101.
- Schulz, A. & Wirtz, M. (2012). Analyse kausaler Zusammenhänge als Ziel des Experimentierens. In W. Rieß, M. Wirtz, B. Barzel & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht* (S. 39–56). Münster: Waxmann.
- Schulz, A., Wirtz, M. & Starauschek, E. (2012). Das Experiment in den Naturwissenschaften. In W. Rieß, M. Wirtz, B. Barzel & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht* (S. 15–38). Münster: Waxmann.
- Shavelson, R. J., Baxter, G. P., Pine, J., Yuré, J., Goldman, S. R. & Smith, B. (1991). Alternative Technologies for Large Scale Science Assessment: Instrument of Education Reform. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 2(2), 97–114.
- Sonnefeld, U. & Kattmann, U. (2002). Lebensräume helfen ordnen: Schülerinnen und Schüler klassifizieren Wirbeltiere. *Zeitschrift für Didaktik der Naturwissenschaften*, 8, 23–31.
- Stecher, B. M., & Klein, S. P. (1997). The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educational Evaluation and Policy Analysis*, 19(1), 1–14.
- Šula, J. (1968). Das Vergleichen und seine Bedeutung für die Bildung elementarer biologischer Begriffe. *Der Biologieunterricht*, 3, 21–39.
- Tomkins, S. P. & Tunnicliffe, S. D. (2001). Looking for ideas: observation, interpretation and hypothesis-making by 12-years-old pupils undertaking science investigations. *International Journal of Science Education*, 23(8), 791–813.
- Trowbridge, J. E. & Mintzes, J. J. (1988). Alternative Conceptions in Animal Classification: a Cross-Age Study. *Journal of Research in Science Teaching*, 25(7), 547–571.
- Walpuski, M., Kampa, N., Kauertz, A. & Wellnitz, N. (2008). Evaluation der Bildungsstandards in den Naturwissenschaften. *Der mathematische und naturwissenschaftliche Unterricht*, 61(6), 323–326.

- Walpuski, M., Kauertz, A., Kampa, N., Fischer, H. E., Mayer, J., Sumfleth, E., et al. (2010). ESNaS – Evaluation der Standards für die Naturwissenschaften in der Sekundarstufe I. In A. Gehrmann, U. Hericks & M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle – Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 171–184). Bad Heilbrunn: Klinkhardt.
- Walpuski, M., Ropohl, M. & Sumfleth, E. (2011). Students' knowledge about chemical reactions – development and analysis of standard-based test items. *Chemistry Education Research and Practice*, 12, 174–183.
- Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung*. Berlin: Logos.
- Wellnitz, N. & Mayer, J. (2012). Beobachten, Vergleichen und Experimentieren: Wege der Erkenntnisgewinnung. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 5* (S. 63–79). Innsbruck: StudienVerlag.
- Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A. et al. (2012). Evaluation der Bildungsstandards – eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften*, 8, 261–291.
- Windschitl, M. (2004). Folk Theories of “Inquiry:” How Preservice Teachers Reproduce the Discourse and Practices of an Atheoretical Scientific Method. *Journal of Research in Science Teaching*, 41(5), 481–512.
- Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *Acer ConQuest. Version 2.0. Generalised Item Response Modelling Software*. Camberwell: Australian Council for Educational Research.

KONTAKT

Nicole Wellnitz
Universität Kassel
Didaktik der Biologie
Heinrich-Plett-Str. 40
34132 Kassel
nicole.wellnitz@uni-kassel.de

AUTORENINFORMATION

Dr. rer. nat. Nicole Wellnitz arbeitet als wissenschaftliche Mitarbeiterin in der Abteilung Didaktik der Biologie an der Universität Kassel. Ihre Forschungsschwerpunkte liegen im Bereich der Diagnose und Entwicklung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung sowie im Bereich motivationaler Einflussfaktoren auf Leistungsunterschiede in Kompetenztests.

Dr. rer. nat. Jürgen Mayer ist Professor für Didaktik der Biologie an der Universität Kassel. Seine Forschungsschwerpunkte sind die Kompetenzmodellierung im Biologieunterricht, die Entwicklung und Analyse von Lernumgebungen, insbesondere des Forschenden Lernens sowie Lehrerprofessionalisierung.