Styles of Scientific Reasoning - Entwicklung eines Testinstruments zur Erfassung prozeduralen und epistemischen Wissens

Nubia Vogt & Dirk Krüger

n.vogt@fu-berlin.de – dirk.krueger@fu-berlin.de

Freie Universität Berlin: Didaktik der Biologie,

Schwendenerstraße 1, 14195 Berlin

Zusammenfassung

Scientific reasoning wird in der Literatur mit verschieden Denkoperationen oder Zielen verknüpft und oft als wissenschaftliches Denken, Schlussfolgern oder Problemlösen aufgefasst. Bisherige Herangehensweisen, das Konstrukt scientific reasoning theoretisch zu strukturieren, setzen häufig einen Fokus auf bestimmte Arbeitsweisen wie das Beobachten, Experimentieren und Modellieren. KIND und OSBORNE (2017) entwickeln diese Strukturierungen weiter, indem sie sechs styles of scientific reasoning beschreiben und für diese neben ontologischen auch prozedurale und epistemische Wissenselemente ausdifferenzieren. Für einige styles of scientific reasoning, vor allem für das Experimentieren, wurden bereits umfassende Untersuchungen von individuellen Denkprozessen durchgeführt, für andere styles fehlen diese jedoch.

Um den theoretischen Vorschlag von KIND und OSBORNE (2017) empirisch überprüfen zu können, werden die styles of scientific reasoning mit Items operationalisiert. Dazu wurden literaturbasiert Kriterien zu den einzelnen styles erarbeitet und darauf aufbauend offene Items entwickelt. Die schriftlichen Antworten von Lehramtsstudierenden auf offene Items zu fünf der sechs untersuchten styles Experimentieren, Modellieren, Kategorisieren und Klassifizieren, probabilistisches Schlussfolgern und historisch-evolutionäres Schlussfolgern dienten als Grundlage für die Entwicklung eines Multiple Choice-Testinstruments. Der Entwicklungsprozess wird mit den Ergebnissen der Pilotierung vorgestellt.

Abstract

In the natural sciences, scientific reasoning is related to different thinking operations or goals in the literature and often defined as scientific thinking or problem solving. Previous approaches to theoretically structure scientific reasoning often focus on scientific methods like observation, experimentation and modeling. KIND und OSBORNE (2017) refine this structuring by describing six styles of scientific reasoning. For each style the authors further characterize besides ontological perspectives procedural and epistemic elements. On few styles of scientific

reasoning, especially experimentation, comprehensive investigations of individual thinking processes have been conducted, for other styles this analysis is missing.

In order to prove the theoretical proposition from KIND und OSBORNE (2017) empirically, the styles of scientific reasoning are operationalized through items. For this purpose criteria for each style have been identified based on the literature. In a second step, open items were developed based on the criteria. The written students' answers to items for the styles experimentation, modelling, categorization and classification, probabilistic reasoning and historic-evolutionary reasoning were then used for the development of a multiple-choice test instrument. The development process together with results from a pilot study are presented in this article.

1 Einleitung

Naturwissenschaftliche Erkenntnisgewinnung ist zentraler Bestandteil einer naturwissenschaftlichen Grundbildung und wird im deutschsprachigen Raum in Denk- (z.B. Hypothesen aufstellen) und Arbeitsweisen (z.B. Experimentieren) unterteilt (DUIT, GROPENGIEßER & STÄUDEL, 2004; NEHRING *et al.*, 2016). Die Anwendung von naturwissenschaftlichen Denk- und Arbeitsweisen spielt eine große Rolle sowohl in der Forschung als auch im naturwissenschaftlichen Unterricht (BYBEE, 2002), weshalb der Erwerb dieser Kompetenzen für Schülerinnen und Schüler von nationalen Bildungsstandards gefordert wird (KMK, 2005). Lernende können durch die Anwendung wissenschaftlicher Denk- und Arbeitsweisen neue Probleme lösen, die Gewinnung naturwissenschaftlicher Erkenntnisse verstehen und mit vorhandenem Fachwissen in Beziehung setzen, was zu einem kritischen und reflektierten Umgang mit naturwissenschaftlichem Wissen führen kann (GROPENGIEßER *et al.*, 2018). MAYER (2007) gliedert in seinem Rahmenkonzept wissenschaftsmethodischer Kompetenzen Erkenntnisgewinnung in die Dimensionen „Wissenschaftliche Arbeitstechniken" (*practical work*), „Wissenschaftliche Untersuchungen" (*scientific inquiry*) und „Charakteristika der Naturwissenschaften" (*nature of science*) und stellt diesen Dimensionen drei Konstrukte zur Seite, die an der Nutzung von Wissen, wie zum Beispiel in Problemlöseprozessen, beteiligt sind. So hängt die Dimension „Wissenschaftliche Untersuchungen" eng mit wissenschaftlichem Denken (*scientific reasoning*) zusammen (KUHN *et al.*, 1988; KLAHR, 2000), welches im Mittelpunkt der Erkenntnisgewinnung steht (HODSON, 2014).

Bis heute fehlt ein Konsens darüber, wie das Konstrukt *scientific reasoning* genau strukturiert werden kann. Eine einheitliche Struktur ist jedoch nötig, um die Kompetenzen in diesem Bereich konkretisieren zu können, damit sie letztendlich diagnostizierbar werden. Ebenso dienen die konkretisierten Kompetenzen zur Erweiterung der Unterrichtsskripte (KMK, 2010). Die Zahl der benannten Elemente für das Konstrukt *scientific resoning* variiert in Abhängigkeit von

Autor, Hintergrund und Fokus der Arbeit stark. In der internationalen Forschung wird *scientific reasoning* beispielsweise häufig mit den Dimensionen *scientific inquiry* und *nature of science* beschrieben, ohne die praktische Dimension miteinzubeziehen (GOTT & DUGGAN, 1995; LEDERMAN, 1992, GIERE *et al.*, 2006). Aufgrund fehlender Klarheit über die Komponenten des wissenschaftlichen Denkens in der Forschung sind die Ziele für die Förderung eines Wissenschaftsverständnisses in der naturwissenschaftlichen Bildung nicht trennscharf definiert. Verschiedene internationale Studien konnten zeigen, dass diese Unklarheit zu einem Fokus auf Fachwissen im Unterricht und in darauffolgenden Kompetenztests führt, da die verschiedenen Strukturierungen von *scientific reasoning* sich nur teilweise überlappen, oft aber auch widersprüchlich sind (LAYTON, 1973; WEIß *et al.*, 2003; DAY & MATTHEWS, 2008). Viele Kompetenztests fordern nur Kompetenzen auf niedrigerem Niveau (z.B. Verständnis) und prüfen keine höheren analytischen und argumentierenden Fähigkeiten (OSBORNE & RATCLIFFE, 2002), was laut KIND und OSBORNE (2017) aus der Unklarheit über das komplexe Konstrukt resultiert. Aufgrund der zahlreichen unterschiedlichen Strukturierungen von *scientific reasoning* besteht eine Vielzahl von Testinstrumenten, die einzelne Komponenten des Konstrukts erfassen (OPITZ, 2017). Es mangelt jedoch sowohl an einer umfassenden Definition des Konstrukts *scientific reasoning* als auch an einem geeigneten Testinstrument, um Kompetenzen von Lernenden in diesem Bereich zu untersuchen.

Ziel dieses Projekts ist es, eine von KIND und OSBORNE (2017) neu vorgeschlagene theoretische Strukturierung von *scientific reasoning* empirisch zu überprüfen und ein Testinstrument bereitzustellen, das alle Wissenselemente, die naturwissenschaftliches Denken ausmachen, gleichermaßen berücksichtigt.

2 Theorie

Bisherige Ansätze, das Konstrukt *scientific reasoning* zu strukturieren, folgen häufig dem hypothetisch-deduktiven Ansatz (POPPER, 1934) und setzen ihren Fokus auf bestimmte Arbeitsweisen (z.B. Beobachten, Experimentieren, Vergleichen, Modellieren; KLAHR & DUNBAR, 1988). GIERE *et al.* (2006) konzentrieren sich auf den Modellierungsprozess als zentralen Aspekt von *scientific reasoning* und argumentieren für ein generelles Muster, das *scientific reasoning* zugrunde liegt. Sie sprechen damit für einen großen Forschungskorpus, der die (Denk-) Prozesse, die *scientific reasoning* ausmachen, zu der „wissenschaftlichen Methode“ verallgemeinert (vgl. MAYER, 2007; SCHWARTZ *et al.*, 2004).

KIND und OSBORNE (2017) kritisieren diese eingeschränkte Sicht auf das Konstrukt und argumentieren für eine Strukturierung von *scientific reasoning*, die sie aus historischer Literatur über die Geschichte der Naturwissenschaften entnommen haben (CROMBIE, 1994). Die Geschichte zeigt, dass es keine singuläre Form des *scientific reasoning* in den Naturwissenschaften je gab oder heute gibt, weshalb die Autoren in ihrem Artikel sechs distinkte Formen von *scientific reasoning* unterscheiden. Die Autoren diskutieren weiterhin, dass vorherige Erklärungen von *scientific reasoning* das Konstrukt als unabhängig vom Fachwissen und Kontext beschreiben, trotz zahlreicher Studien, die eine Kontextabhängigkeit zeigen konnten (z.B. GIERE *et al.*, 2006; ZIMMERMANN, 2007). Ausgehend von dieser Forschung plädieren KIND und OSBORNE (2017) für eine umfassendere Betrachtung von *scientific reasoning*-Kompetenzen in einem fachwissenschaftlichen Kontext.

2.1 Styles of scientific reasoning

KIND und OSBORNE (2017) schlagen in ihrem Artikel basierend auf der Geschichte der Naturwissenschaft (CROMBIE, 1994) eine Strukturierung von *scientific reasoning* in sechs distinkte *styles of scientific reasoning* vor, die kontextspezifisch unterschiedlich häufig vorkommen und verschieden stark ausgeprägt sein können. Die Autoren stellen die sechs *styles* der eindimensionalen Strukturierung von *scientific reasoning* gegenüber und argumentieren zusätzlich für eine Abhängigkeit der einzelnen *styles* von drei Wissensdimensionen. Die Wissensdimensionen ergeben sich aus dem Ziel des *scientific reasoning*, drei Fragen über die belebte Welt zu beantworten: 1. Was existiert?, 2. Warum geschieht es?, 3. Woher wissen wir das? (OSBORNE, 2011). Diese Wissensdimensionen sind zum einen das Kontextwissen (ontologisches Wissen) über die Inhalte der Naturwissenschaft, zum anderen das prozedurale Wissen über die Prozesse in der Naturwissenschaft und die kognitiven Werkzeuge die notwendig sind, um den jeweiligen *style* ausführen zu können, und zuletzt das übergeordnete epistemische Wissen über Konstrukte und Methoden, die essentiell für den Prozess der Wissensbildung in der Wissenschaft sind (OECD, 2017; ARNOLD *et al.*, 2014).

Die von KIND und OSBORNE (2017) vorgeschlagenen *styles of scientific reasoning* gliedern sich in die naturwissenschaftlichen Arbeitsweisen *(i) Experimentieren, (ii)Modellieren* sowie *(iii) Kategorisieren und Klassifizieren* und die Denkweisen *(iv) probabilistisches Schlussfolgern, (v) historisch-evolutionäres Schlussfolgern* und *(vi) mathematische Deduktion*. Das *Experimentieren* dient vor allem der Überprüfung von Hypothesen und der Schlussfolgerung von kausalen Zusammenhängen, kann aber auch induktiv zur Entwicklung einer verallgemeinernden Gesetzmäßigkeit genutzt werden. Zentrale Aspekte sind

hierbei beispielsweise die Variablenkontrolle oder die angemessene Interpretation der gewonnenen Daten (WELLNITZ & MAYER, 2013). Mit Hilfe des *Modellierens* lassen sich ebenfalls Hypothesen testen, das Modell kann auf Basis der gewonnenen Erkenntnisse auch angepasst werden. Zu reflektierende Aspekte sind unter anderem die Eigenschaften oder der Zweck von Modellen (KRELL, UPMEIER ZU BELZEN & KRÜGER, 2016; UPMEIER ZU BELZEN & KRÜGER, 2010). Das *Kategorisieren und Klassifizieren* umfasst alle Aspekte, die nötig sind, um die Vielfalt in einem naturwissenschaftlichen Gebiet ordnen zu können. Dabei spielen vor allem das Vergleichen und das Finden zweckmäßiger Kriterien eine tragende Rolle, zum Beispiel im Kontext der Systematik (JANICH & WEINGARTEN, 1999; HAMMANN, 2002). Das *probabilistische Schlussfolgern* ist entstanden aus dem Bedarf einer präzisen Logik, um Entscheidungen in unsicheren Situationen treffen zu können, und enthält statistisches Wissen über Wahrscheinlichkeiten, Testverfahren und Diagramme (PFANNKUCH & WILD, 2004; BEN-ZVI *et al.*, 2018). Das *historisch-evolutionäre Schlussfolgern* befasst sich mit vergangenen Ereignissen und deren Entstehung und versucht, mit Hilfe rückwärtsgerichteter Hypothesen, die einer abduktiven Lösungsfindung entsprechen, die Ursachen zu erklären (MAYR, 2004; SÜßMUTH, 2007; JUNKER & SCHERER, 2001). Der *style mathematische Deduktion* bildet ein übergeordnetes Denkmuster, die Deduktion, in scharfer Abgrenzung zur Induktion und Abduktion ab. KIND und OSBORNE (2017) leiten ihn nach CROMBIES (1994) Vorlage aus der Historie ab und beschreiben in dem *style* die mathematischen Anfänge der naturwissenschaftlichen Forschung. Als prozedurale Elemente nennen sie mathematische Verfahren wie die Integration und Geometrie. Da das Testinstrument kontextabhängig (biologiespezifisch) entwickelt werden soll und der *style mathematische Deduktion* als Denkweise in jedem Verfahren beim Generieren von Hypothesen aus einer Theorie enthalten ist (POPPER, 1934), wird er für die folgende Fragestellung nicht weiter separat betrachtet.

3 Fragestellung

Um den theoretischen Vorschlag von KIND und OSBORNE (2017) empirisch überprüfen zu können, werden die *styles of scientific reasoning* mit Items operationalisiert, die das theoretische Modell in Form empirisch erfassbarer Indikatoren abbilden. Daraus ergibt sich folgende Forschungsfrage:

F1: Inwiefern lässt sich die vorgeschlagene Strukturierung von *scientific reasoning* in fünf unterschiedliche *styles* mit ihren jeweiligen prozeduralen und epistemischen Wissenselementen empirisch abbilden?

Um dieser Frage nachzugehen, wird ein Multiple-choice Testinstrument mit biologiespezifischen Items zu fünf *styles* entwickelt. Es werden separate Items zu prozeduralen sowie epistemischen Wissenselementen für die verbleibenden Bereiche konstruiert.

4 Methode

Der Konstruktionsprozess für das zu entwickelnde Testinstrument orientiert sich an Vorschlägen von TERZER *et al.* (2013) und MATHESIUS *et al.* (2014) (Abb. 1).

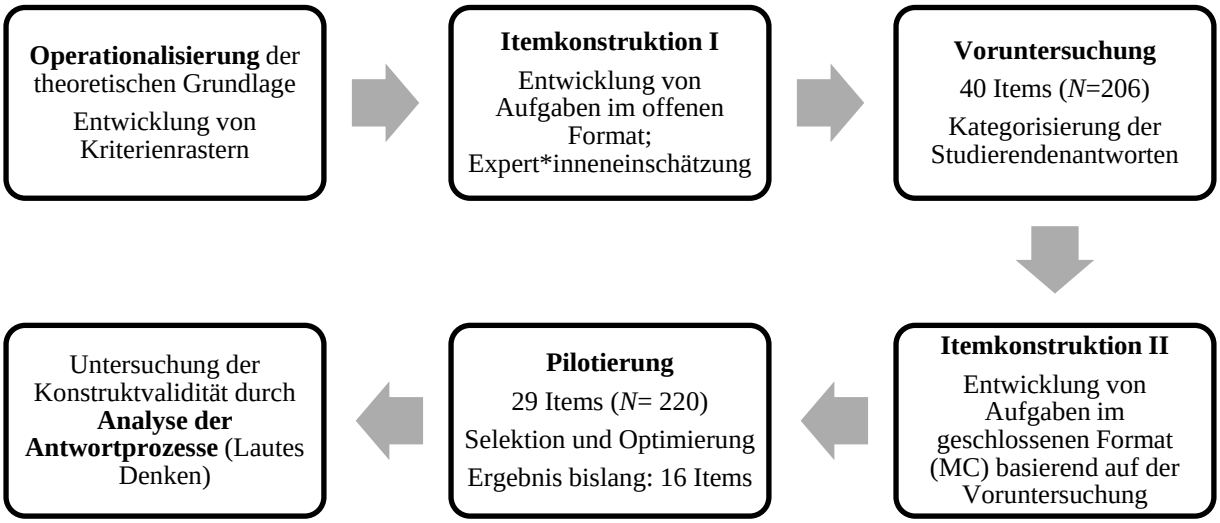


Abbildung 1: Prozess der Testentwicklung (verändert nach MATHESIUS *et al.*, 2014).

In sechs Arbeitsschritten soll ein paper-pencil-Testinstrument mit geschlossenen Multiple-choice (MC) Items entstehen, die im Vergleich zu Items im offenen Format eine gesteigerte Testökonomie sowie eine erhöhte Auswertungsobjektivität garantieren (MOOSBRUGGER & KELAVA, 2012). Zusätzlich ermöglicht das MC-Format die Beantwortung einer höheren Itemanzahl pro getesteter Person innerhalb des Bearbeitungszeitraums. Ausgehend von RODRIGUEZ‘ (2005) Ergebnissen wurden in diesem Projekt Items mit drei Antwortoptionen (Ein Attraktor, zwei Distraktoren) konstruiert. Rodriguez konnte in seiner Literaturanalyse zeigen, dass drei Antwortoptionen im Vergleich zu vier Antwortoptionen die Itemschwierigkeit senken und die Reliabilität erhöhen. Die Konstruktion von nur drei Antwortoptionen zieht außerdem praktische Vorteile nach sich. Es wird weniger Zeit für die Itemkonstruktion benötigt, die Bearbeitungszeit der einzelnen Items sinkt, sodass eine noch höhere Anzahl beantwortet werden kann, und es werden weniger mögliche Hinweise zur Beantwortung der Frage gegeben.

Im ersten Schritt wurden basierend auf dem Vorschlag von KIND und OSBORNE (2017) literaturgestützt biologiespezifische Kriterienraster für die *styles (i) Experimentieren, (ii) Modellieren, (iii) Kategorisieren und Klassifizieren, (iv) probabilistisches Schlussfolgern* und *(v) historisch-evolutionäres Schlussfolgern* erstellt. Innerhalb der einzelnen *styles* wurde zwischen den Teildimensionen prozedural und epistemisch unterschieden. Auf die Betrachtung der ontologischen Wissensdimension wurde verzichtet, da Denkstrukturen erfasst werden sollen, die über die Wiedergabe von Fachwissen hinausgehen, die in zahlreichen Kompetenztests geprüft wird (Kap. 1). Die Kriterienraster dienen als Basis zur Operationalisierung des theoretischen Vorschlags einer Strukturierung von *scientific reasoning* und sind Grundlage für die Entwicklung von Items, mit deren Hilfe die verschiedenen Ausprägungen des Konstrukts empirisch abgebildet werden sollen (BORTZ, 1984).

Im Schritt Itemkonstruktion I (Abb. 1) wurden Items im offenen Format zu den einzelnen Kriterien der untersuchten *styles* entwickelt (N=40). Dieser Schritt wurde unternommen, um auf Basis der Antworten authentische MC-Antwortoptionen zu entwickeln. Der Kontext der Itemstämme ist, unabhängig vom *style*, immer ein biologischer, und kann einfache Abbildungen, Graphen oder Tabellen enthalten, um die Itemschwierigkeit zu senken (PRENZEL *et al.*, 2002). Die fachliche Korrektheit und die Passung der Items zu den Kriterien wurde durch Expert*innen aus der Fachwissenschaft (Systematik, Evolution, Statistik) und der Biologiedidaktik, die ein hohes Level an theoretischem Wissen und Forschungserfahrung aufweisen, geprüft und diskutiert.

Die Items im offenen Format wurden in einer Voruntersuchung bei N=206 Lehramtsstudierenden mit Biologie als Haupt- oder Nebenfach an der Freien Universität Berlin eingesetzt. Die Antworten wurden sortiert und auf ihre fachliche Korrektheit hin beurteilt.

Ausgehend von den erhobenen Studierendenaussagen wurden drei MC-Antwortoptionen entwickelt (Itemkonstruktion II; Abb. 1), die den vorgeschlagenen Richtlinien zur Konstruktion von Testaufgaben von JONKISZ, MOOSBRUGGER und BRANDT (2012) und HALADYNA (2004) folgen. So wurde unter anderem darauf geachtet, dass die Antwortoptionen möglichst keine Fremdwörter enthalten und in etwa gleich lang und ähnlich syntaktisch aufgebaut sind.

Die entstandenen MC-Items (N=29) wurden in einer Pilotierung (Abb. 1) im Zeitraum vom Sommersemester 2019 bis Wintersemester 2019/2020 bei N = 220 Studierenden an der Freien Universität Berlin und der Humboldt-Universität zu Berlin eingesetzt. Die Verteilung der Items auf die zu erfassenden Facetten ist in Tabelle 4 dargestellt. Um die Eignung der MC-Items zu beurteilen, wurde die

Itemschwierigkeit auf Basis der Lösungswahrscheinlichkeit berechnet. Um später eine angemessene diagnostische Funktion zu gewährleisten, werden zu einfache und zu schwierige Items ausgeschlossen. Die Items sollen eine möglichst breite Schwierigkeitsstreuung im Bereich zwischen 0,2 und 0,8 aufweisen, um Personen mit unterschiedlichen Fähigkeiten gleich gut differenzieren zu können (BORTZ & DÖRING, 2006). Für MC-Items kann in einer Distraktorenanalyse anhand von erhobenen Daten zusätzlich untersucht werden, wie häufig die einzelnen Distraktoren gewählt wurden (LIENERT & RAATZ, 1998). Distraktoren, die sehr häufig oder sehr selten gewählt wurden, wurden überarbeitet.

5 Ergebnis

In den folgenden Abschnitten wird der Prozess der Itementwicklung an einem Item des *styles Kategorisieren und Klassifizieren* beispielhaft erläutert, das bereits nach der ersten Pilotierung eine optimale Verteilung der Antworten aufwies.

5.1 Kriterienraster als Basis der Operationalisierung

Die Kriterien für die Kriterienraster wurden aus KIND und OSBORNES (2017) Artikel entnommen und durch Aspekte aus einschlägiger Fachliteratur und fachdidaktischer Literatur ergänzt (Kap. 2.1). Für den *style Kategorisieren und Klassifizieren* wurde vor allem Literatur zu den Themen Wissenschaftstheorie und Systematik genutzt (Tab. 1). Die Vollständigkeit der Kriterienraster wurde mit Fachdidaktiker*innen diskutiert. Nach der Diskussion wurden keine weiteren Aspekte hinzugefügt und die Kriterienraster schienen das Konstrukt angemessen zu repräsentieren. Es ergeben sich je drei prozedurale und epistemische Kriterien im *style Kategorisieren und Klassifizieren*.

Tabelle 1: Kriterienraster für den *style Kategorisieren und Klassifizieren*.

prozedural	**epistemisch**
Studierende können ..	Studierende wissen, dass ..
(1) .. Ordnen als Methode anwenden (DUIT *et al.*, 2004).	(1) .. Ordnungssysteme durch neue Erkenntnisse veränderbar sind (JANICH *et al.*, 2001).
(2) .. kriterienstetes Vergleichen als Methode anwenden (HAMMANN, 2002).	(2) .. es Grenzen bei der Erstellung von Ordnungssystemen gibt: a) nicht trennscharfe Klassen b) Restgrößen (STÄUDEL *et al.*, 2002).
(3) .. kontextabhängige Kriterien zum Ordnen & Vergleichen identifizieren (KWA & MCKAY, 2011).	(3) .. das Vergleichen eine dreistellige Funktion ist (JANICH &WEINGARTEN, 1999).

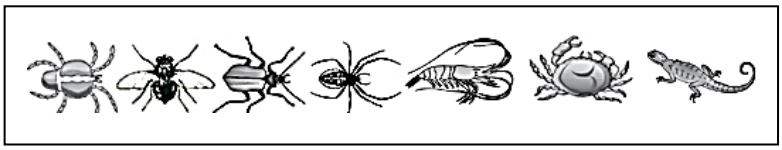
5.2 Itemkonstruktion I & Voruntersuchung

Auf Basis der gefundenen Kriterien wurden zuerst Items im offenen Format entwickelt. Die Kontexte sind biologiespezifisch und an den Forschungsalltag angelehnt. Wenn nötig, wurden die Itemstämme durch passende Abbildungen, Graphen oder Tabellen ergänzt. Die Items (N=40) wurden auf mehrere Testhefte verteilt und von Lehramtsstudierenden (N=206) schriftlich beantwortet. Die Antworten der Studierenden wurden auf ihre fachliche Korrektheit hin geprüft und sortiert. Antworten, die Alltagsvorstellungen enthielten, wurden markiert und dienten als Grundlage für die Formulierung glaubwürdiger Distraktoren für die Itemkonstruktion II (Abb.1). SADLER (1998) konnte zeigen, dass Distraktoren, die auf Grundlage von Antworten auf offene Impulse entstanden sind, häufiger gewählt werden als frei konstruierte Distraktoren.

5.3 Itemkonstruktion II & Pilotierung

Die entwickelten MC-Items (N=29, davon N=6 zum *style Kategorisieren und Klassifizieren*) wurden in einer Pilotierung eingesetzt und nach Prüfung der Itemschwierigkeit und einer Distraktorenanalyse gegebenenfalls optimiert. Dafür wurden alternative Studierendenantworten genutzt oder die Formulierungen der Antwortoptionen angepasst. In einer weiteren Pilotierung (N=220) wurden die überarbeiteten Items erneut geprüft.

Tabelle 2: Selektiertes Item KK_{pro1} (Tab. 3) für das prozedurale Kriterium (1) (Tab. 1) mit ausgewählten Studierendenantworten und daraus abgeleiteten Antwortoptionen; p_i = Itemschwierigkeit.

Item	Offenes Antwortformat	Geschlossenes Antwortformat
Stamm	Eine Biologielehrerin teilt in einer Unterrichtsstunde Abbildungen verschiedener Lebewesen aus und erteilt den Arbeitsauftrag, diese zu klassifizieren. Ein Schüler präsentiert als Ergebnis folgende Anordnung: klein  groß **Abbildung:** Anordnung verschiedener Lebewesen.	
Impuls	*Beschreiben und beurteilen Sie das Ergebnis in Bezug auf den Arbeitsauftrag.*	*Entspricht das präsentierte Ergebnis des Schülers einer wissenschaftlichen Klassifikation?*
Antworten	Der Schüler versteht anscheinend nicht die Bedeutung des Wortes "klassifizieren". Er hat, statt die Lebewesen in Klassen einzuordnen, sie einfach anhand eines Merkmals (der Größe) in eine Reihenfolge gebracht.	Nein, da der Schüler die Tiere nach ihrer Größe ordnet. [*Attraktor; n=21 ≙ 45%*]
	Der Schüler hat die Tiere nach Größe geordnet und hätte sie taxonomisch vielleicht wissenschaftlicher sortiert. [...] bleibt er bei einer Eigenschaft, nach der er ordnet: Kriterienstet vs. kriterienunstet.	Ja, da der Schüler die Tiere kriterienstet sortiert. [*Distraktor 1; n=10 ≙ 21%*]
	Der Schüler wählt ein Kriterium, welches er aus dem Alltag kennt. Er scheint nicht das wissenschaftliche Vorgehen zu kennen bzw. kennt die Kriterien nicht, die in der Wissenschaft für Klassifikationen angewendet werden.	Die Frage kann nicht begründet beantwortet werden, da der Schüler kein Klassifikationskriterium wählt. [*Distraktor 2; n=16 ≙ 34%*]
Itemwerte	$p_i = 0{,}45$ | $N = 47$	

5.4 Ergebnis der Pilotierung

Nach Analyse und Optimierung der entwickelten MC-Items (N=29) wurden bislang N=16 Items selektiert. Die Itemschwierigkeit (p_i) wird berechnet, indem die Anzahl richtiger Lösungen durch die Gesamtzahl der Lösungen geteilt wird. Je höher die Zahl, desto leichter ist das Item. Die Itemschwierigkeit liegt nach der ersten Optimierung zwischen 0,19 und 0,78 (M = 0,44). Die Distraktoren der selektierten optimierten Items wurden von mindestens 9 % und höchstens 62 % der Studierenden gewählt (M = 28,3 %) (Tab. 3).

Tabelle 3: Ergebnisse der Distraktorenanalyse für die selektierten Items vor und nach der Optimierung. Die Itemschwierigkeit (p_i) entspricht dem Attraktor/100. Für Items, bei denen keine Optimierung nötig war, liegen keine Werte „Nach der Optimierung" vor. E = *Experimentieren*; M = *Modellieren*; KK = *Kategorisieren und Klassifizieren*; PS = *Probabilistisches Schlussfolgern*; HS = *Historisch-evolutionäres Schlussfolgern*; X_p = prozedural; X_e = epistemisch; A = Attraktor, $D_{1/2}$ = Distraktoren.

	Vor Optimierung			**Nach Optimierung**		
Items	A/p_i	D_1	D_2	A/p_i	D_1	D_2
E_{e1}	32 %	12 %	56 %	-	-	-
E_{e2}	19 %	19 %	62 %	-	-	-
E_{e3}	29 %	33 %	38 %	-	-	-
M_{e1}	41 %	26 %	33 %	-	-	-
M_{e3}	44 %	16 %	40 %	-	-	-
KK_{p1}	45 %	21 %	34 %	-	-	-
KK_{p4}	76 %	12 %	12 %	-	-	-
KK_{e1}	54 %	0 %	46 %	24 %	20 %	56 %
KK_{e2}	31 %	32 %	37 %	-	-	-
PS_{p2}	78 %	0%	22 %	58 %	13 %	29 %
PS_{e1}	39 %	20 %	41 %	-	-	-
PS_{e2}	40 %	8 %	52 %	48 %	9 %	43 %
PS_{e3}	82 %	9 %	9 %	69 %	12 %	19 %
PS_{e4}	54 %	4 %	42 %	40 %	15 %	45 %
HS_{e2}	69 %	6 %	25 %	52 %	13 %	35 %
HS_{e3}	48 %	25 %	27 %	-	-	-

Weitere *N*=10 Items befinden sich zurzeit noch in der Optimierung, da sie als zu einfach oder schwierig eingeschätzt werden. *N*=3 Items wurden von der Studie ausgeschlossen, da sie nicht optimierbar waren. Jede Facette wird mit höchstens vier Items erfasst.

Tabelle 4: Verteilung der *N*=26 Items auf die zu erfassenden Facetten des *scientific reasoning*, getrennt nach bereits selektierten und zu optimierenden Items. Die *N*=3 von der Studie ausgeschlossenen Items werden nicht aufgeführt.

style	**prozedural**		**epistemisch**		**gesamt**	
	selektiert	Optimierung	selektiert	Optimierung	selektiert	Optimierung
Experimentieren	-	-	**3**	2	**3**	2
Modellieren	-	-	**2**	1	**2**	1
Kategorisieren & Klassifizieren	**2**	1	**2**	1	**4**	2
Probabilistisches Schlussfolgern	**1**	1	**4**	-	**5**	1
Historisch-evolutionäres Schlussfolgern	-	1	**2**	3	**2**	4
gesamt	**3**	3	**13**	7	**16**	10

6 Diskussion

KIND und OSBORNE (2017) liefern mit ihrer Strukturierung von *scientific reasoning* in sechs distinkte *styles of scientific reasoning* einen historisch begründeten Vorschlag, welche Denk- und Arbeitsweisen mit den jeweils zugehörigen prozeduralen und epistemischen Wissenselementen Dimensionen von *scientific reasoning* sein können. Aufgrund der unterschiedlichen Auffassungen der Dimensionen und Dimensionalität von *scientific reasoning* (z.B. MAYER, 2007; LEDERMAN, 2002; FISCHER *et al*., 2014) gibt es kein Standard-Testinstrument, sondern zahlreiche verschiedene, die ihren Fokus auf einzelne Denk- und Arbeitsweisen legen oder nicht zwischen ontologischen, prozeduralen und epistemischen Wissenselementen unterscheiden (OPITZ, 2017). OPITZ (2017) schließt aus seinen Analysen von 38 *scientific reasoning*-Testinstrumenten, dass es einen Trend hin zur Konzeptualisierung von *scientific reasoning* als domänenspezifisches Set verschiedenartiger, aber zusammenhängender Wissenselemente gibt. Das hier vorgestellte Testinstrument (Kap. 4 und 5) knüpft an OPITZ‘ (2017) Ergebnisse an, indem es domänenspezifische Items zu fünf *styles of scientific reasoning* bereitstellt, mit Hilfe derer zukünftig Kompetenzen im Bereich *scientific reasoning* bei Lehramtsstudierenden der Biologie diagnostiziert werden können. Auf die Entwicklung von Items zum sechsten *style*, der *mathematischen Deduktion*, wurde verzichtet, da dieser keine Items in einem biologiespezifischen Kontext hervorbringt und als übergeordnete Denkweise Bestandteil in anderen *styles* wie dem *Experimentieren* und *Modellieren* ist, oder anderen Denkweisen gegenübergestellt wird, insbesondere im *style historisch-evolutionäres Schlussfolgern*, in dem die Abduktion als vorrangige Denkweise genutzt wird (KIND & OSBORNE, 2017; MAYR, 2004). Das vorgestellte Testinstrument erfasst durch die Strukturierung des Konstrukts *scientific reasoning* in verschiedene *styles* und Wissenselemente (prozedural, epistemisch) kontextabhängig mehr Facetten als bestehende Testinstrumente, die typische Kompetenzen wie das Generieren von Hypothesen und die Auswertung von Daten erfassen (OPITZ, 2017). Mit dem schriftlichen Testinstrument können kognitive Kompetenzen im Bereich *scientific reasoning* erfasst werden, fraglich bleibt jedoch, inwieweit damit echte Problemlösekompetenz und nicht methodisches Wissen erhoben wird (MAYER, 2007). Mit Blick auf die Verteilung der Items auf die einzelnen Facetten des Konstrukts lässt sich feststellen, dass bisher deutlich mehr Items für die epistemischen Facetten entwickelt und optimiert werden konnten. Für die *styles Experimentieren* und *Modellieren* wurden nach Diskussion in der Arbeitsgruppe alle prozeduralen Items ausgeschlossen, da diese Facetten mit einem MC-Testinstrument nicht erfassbar scheinen. Im Hinblick auf die absolute Anzahl der Items wäre zu überdenken, die prozeduralen Items der anderen *styles* ebenfalls

auszuschließen und ein rein epistemisches Testinstrument zu entwickeln. Die Kontexte der Items sind austauschbar, während das zugrunde liegende Kriterium erhalten bleibt. Alternative Items zu den *N*=29 vorgestellten Items sind in Planung, um den Itempool zu erweitern und eine genügende inhaltliche Abdeckung der verschiedenen Facetten sicherzustellen.

Während der Testentwicklung wurden verschiedene Maßnahmen berücksichtigt, um Evidenz für die Validität der Testwertinterpretationen zu gewinnen (AREA *et al.*, 2004; SCHMIEMANN & LÜCKEN, 2014). Die Inhaltsvalidität wurde überprüft, indem Expert*innen der aus Biologiedidaktik, die sich im Rahmen ihrer Forschung mit dem Konstrukt *scientific reasoning* beschäftigen, einschätzten, inwieweit die Items das Konstrukt angemessen abbilden. In einer Voruntersuchung wurden die Studierenden durch das Anbieten der Items mit einem standardisierten offenen Impuls zu einer eigenständigen Antwortfindung angeregt. Die schriftlichen, unter fachlicher Perspektive breit streuenden Antworten konnten als Grundlage für die Entwicklung der MC-Antwortoptionen genutzt werden. Die Entwicklung von drei Antwortoptionen bot die Möglichkeit, nur häufig vertretene fachlich inkorrekte Vorstellungen der Studierenden als Distraktoren umzusetzen (RODRIGUEZ, 2005). Die Antworten gaben außerdem Hinweise auf mögliche Verständnisprobleme des Itemstamms, der daraufhin bei einigen Items überarbeitet wurde. Es war möglich, früh im Entwicklungsprozess geeignete Itemstämme zu identifizieren. Die Analyse der Itemkennwerte für die konstruierten MC-Aufgaben, wie hier die Häufigkeit der Distraktorenwahl und die Berechnung der Itemschwierigkeit, erlaubte zusätzlich Rückschlüsse auf die Güte des Testinstruments. Es lässt sich erkennen, dass alle konstruierten Antwortoptionen bei den selektierten Items attraktiv sind, im Durchschnitt für etwa ein Drittel der Studierenden, sodass ihnen eine gewisse Plausibilität unterstellt werden kann (SADLER, 1998). Insgesamt wurden bislang 16 Items aufgrund ihrer Kennwerte ausgewählt. Bei 10 von 29 Items konnte bereits nach der ersten Pilotierung eine Itemschwierigkeit berechnet werden, die sich im Rahmen der von BORTZ und DÖRING (2006) vorgeschlagenen Richtwerte befindet. Sie zeigten außerdem eine angemessene Verteilung der Antworten auf die Antwortoptionen und mussten nicht überarbeitet werden. Nach Optimierung der restlichen Items konnte weitere sechs Items in den Itempool übernommen werden. Die übrigen zehn befinden sich in einer weiteren Optimierungsphase, zusätzlich sind weitere Items mit alternativen Kontexten in Planung.

7 Ausblick

Es konnte gezeigt werden, dass die Entwicklung eines MC-Testinstruments auf Basis von erhobenen Studierendenantworten zur validen Interpretation der Testergebnisse beiträgt (MESSICK, 1995; AERA *et al.*, 2004). In einer folgenden Testphase wird Evidenz für Validität basierend auf Antwortprozessen gesucht, indem Lehramtsstudierende verschiedener Fächerkombinationen mittels lauten Denkens (ERICSSON & SIMON, 1980) dazu aufgefordert werden, ihre Lösungsprozesse bei der Bearbeitung der Items zu verbalisieren. Die Lösungsprozesse geben Aufschluss darüber, ob die kognitiven Prozesse zur Lösung der Items zu den vorher erstellten Kriterien passen, oder ob nicht beabsichtigte Prozesse wie Raten oder Unsicherheit in Bezug auf kontextspezifische Begriffe zur Lösung führen (HARTIG *et al.*, 2012). Die verschriftlichten Protokolle der verbalen Daten werden kategoriengeleitet ausgewertet und können dann mit Hilfe statistischer Verfahren weiter analysiert werden (SANDMANN, 2014). Ferner werden die Erkenntnisse genutzt, um die Items weiter zu optimieren.

Die Analysen auf Itemebene sollen zusätzlich durch Analysen der internen Struktur auf Testebene erweitert werden, um zu untersuchen, inwieweit die vorgeschlagene Strukturierung von KIND und OSBORNE (2017) von *scientific reasoning* in distinkte *styles* tragfähig ist. Hinweise auf ein mehrdimensionales Konstrukt würden die Annahme der Autoren stützen, dass *scientific reasoning* in voneinander abgrenzbare Denk- und Arbeitsweisen strukturiert werden kann.

Zitierte Literatur

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION [AERA, APA & NCME] (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

ARNOLD, J., KREMER, K., & MAYER, J. (2014). Understanding students' experiments – What kind of support do they need in inquiry tasks? *International Journal of Science Education*, 36 (15–16), 2719-2749.

BEN-ZVI, D., MAKAR, K., & GARFIELD, J. (2018). *International handbook of research in statistics education.* Cham: Springer. https://doi.org/10.1007/978-3-319-66195-7.

BORTZ J. (1984): Vom vorwissenschaftlichen Probleminteresse zur empirischen Untersuchung. In: Lehrbuch der empirischen Forschung. Springer, Berlin, Heidelberg.

BORTZ J., DÖRING, N. (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Springer, Berlin, Heidelberg.

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

CROMBIE, A. C. (1994). *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts*. London, England: Duckworth.

DUIT, R., GROPENGIEßER, H. & STÄUDEL, L. (2004): *Naturwissenschaftliches Arbeiten.* In: *Unterricht und Material 5-10*. Erhard Friedrich Verlag, Seelze-Velber.

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

FISCHER, F., KOLLARA, I., UFERB, S., SODIANA, B., HUSSMANNC, H., PEKRUNA, R., EBERLEA, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 5,* 28-45.

GIERE, R., BICKLE, J. & MAUDLIN, R. F. (2006). *Understanding scientific reasoning.* (5. Auflage). Belmont, CA: Thomson Wadsworth.

GOTT, R., DUGGAN, S. (1995). *Investigative Work in Science Curriculum*. Open Univ. Press, Buckingham.

GROPENGIEßER H., HARMS, U., KATTMANN, U. (2018). *Fachdidaktik Biologie*. Aulis Verlag, Hallbergmoos.

HALADYNA, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (3. Auflage). Routledge, New York.

HAMMANN, M. (2002). *Kriteriengeleitetes Vergleichen im Biologieunterricht.* Studienverlag, Innsbruck.

HARTIG, J., FREY, A., JUDE, N. (2012). *Validität*. In: MOOSBRUGGER, H., KELAVA, A. [Hrsg.]: *Testtheorie und Fragebogenkonstruktion.* Springer-Lehrbuch. Springer, Berlin, Heidelberg.

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

JANICH, R., GUTMANN, M. & PRIEß, K. (2001). *Biodiversität: Wissenschaftliche Grundlagen und gesetzliche Relevanz.* https://doi.org/10.1007/978-3-642-56739-1.

JANICH, P., WEINGARTEN, M. (1999). *Wissenschaftstheorie der Biologie*. UTB, Stuttgart.

JONKISZ, E., MOOSBRUGGER, H., BRANDT, H. (2012). *Planung und Entwicklung von Tests und Fragebogen.* In: MOOSBRUGGER, H., KELAVA, A. [Hrsg.]: *Testtheorie und Fragebogenkonstruktion.* Springer-Lehrbuch. Springer, Berlin, Heidelberg.

JUNKER, R., SCHERER, S. (2001). *Evolution. Ein kritisches Lehrbuch*. Weyel, Gießen.

KIND, P., OSBORNE, J. (2017). Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Science Education, 101,* 8-31.

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

KLAHR, D., DUNBAR, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science, 12*, 1-48.

KMK. (2005). Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004. München.

KMK. (2010). Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung. Köln: Wolters Kluwer.

KRELL, M., UPMEIER ZU BELZEN, A., KRÜGER, D. (2016). *Modellkompetenz im Biologieunterricht.* In: A. SANDMANN, SCHMIEMANN, P. [Hrsg.]: *Biologiedidaktische Forschung: Schwerpunkte und Forschungsstände.* Logos, Band 1, 83-102.

KUHN, D., AMSEL, E., O'LOUGHLIN, M. (1988). *The development of scientific thinking skills*. Academic Press, Orlando.

KWA, C., MCKAY, D. (2011). *Styles of Knowing: A New History of Science from Ancient Times to the Present*. University of Pittsburgh Press, Pittsburgh.

LAYTON, D. (1973). *Science for the people: The origins of the school science curriculum in England.* Allen and Unwin, London.

LEDERMAN, N. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching, 29(4),* 331-359.

LIENERT, G. A., RAATZ, U. (1998). *Testaufbau und Testanalyse*. Beltz, Weinheim.

MATHESIUS, S., UPMEIER ZU BELZEN, A., KRÜGER, D. (2014). Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. Entwicklung eines Testinstruments. *Erkenntnisweg Biologiedidaktik, 13,* 73-88.

MAYER, J. (2007) *Erkenntnisgewinnung als wissenschaftliches Problemlösen*. In: KRÜGER, D., VOGT, H. [Hrsg.]: *Theorien in der biologiedidaktischen Forschung*. Springer-Lehrbuch. Springer, Berlin, Heidelberg.

MAYR, E. (2004). *What makes biology unique? Considerations on the autonomy of a scientific discipline.* Cambridge University Press, Cambridge.

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: MOOSBRUGGER, H. & KELAVA, A. [Hrsg.]: *Testtheorie und Fragebogenkonstruktion*. Springer, Heidelberg. S. 7-26.

NEHRING, A., STILLER, J., NOWAK, K., UPMEIER ZU BELZEN, A., TIEMANN, R. (2016): Naturwissenschaftliche Denk- und Arbeitsweisen im Chemieunterricht – eine modellbasierte Videostudie zu Lerngelegenheiten für den Kompetenzbereich der Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften, 22,* 77-96.

OECD. (2017). PISA 2015 *Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. PISA, OECD Publishing, Paris.

OPITZ, A., HEENE, M., FISCHER, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation, 23(3–4),* 78-101.

OSBORNE, J. F., & RATCLIFFE, M. (2002). Developing effective methods of assessing Ideas and evidence. *School Science Review, 83 (305),* 113-123.

OSBORNE, J. F. (2011). Science teaching methods: A rationale for practices. *School Science Review, 93(343),* 93 – 103.

PFANNKUCH, M., WILD, C. (2004). *Towards an understanding of statistical thinking*. In: BEN-ZVI, D., GARFIELD, J. [Hrsg.]: *The challenge of developing statistical literacy, reasoning, and thinking.* Kluwer Academic Publishers, Dordrecht. S. 17-46.

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

PRENZEL, M., HÄUßLER, P., ROST, J., SENKBEIL, M. (2002). *Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen?*. In: *Unterrichtswissenschaft 30 (2),* S. 120-135.

RODRIGUEZ, M. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice. 24,* 3-13.

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.

SANDMANN, A. (2014). *Lautes Denken – die Analyse von Denk-, Lern- und Problemlöseprozessen*. In: KRÜGER, D., PARCHMANN, I., SCHECKER, H. [Hrsg.]: *Methoden in der naturwissenschaftsdidaktischen Forschung.* Springer-Lehrbuch. Springer, Berlin, Heidelberg. S. 179-188.

SCHMIEMANN, P., LÜCKEN, M. (2014). *Validität - Misst mein Test, was er soll?* In: KRÜGER, D., PARCHMANN, I., SCHECKER, H. [Hrsg.]: *Methoden in der naturwissenschaftsdidaktischen Forschung.* Springer-Lehrbuch. Springer, Berlin, Heidelberg. S. 107-118.

SCHWARTZ, R., LEDERMAN, N., CRAWFORD, B. (2004). Developing Views of Nature of Science in an Authentic Context: An Explicit Approach to Bridging the Gap Between Nature of Science and Scientific Inquiry. *Science Education, 88(4),* 610-645.

STÄUDEL, L., WERBER, B., FREIMANN, T. (2002). *Lernbox Naturwissenschaften: Verstehen und anwenden.* Freidrich Verlag, Seelze-Velber.

SÜßMUTH, R. (2007). Die Evolutionstheorie, ihre Bedeutung und ihre Grenzen. *Imago Hominis 14(1),* 13-45.

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, 15,* 41-57.

WEISS, I. R., PASLEY, J. D., SEAN SMITH, P., BANILOWER, E. R., HECK, D. J. (2003). *A study of K– 12 mathematics and science education in the United States.* Horizon Research, Chapel Hill.

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

ZIMMERMANN, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27(2),* 172-223.