

PHILIPP SCHMIEMANN

Fachsprache in biologischen Testaufgaben

Terminology in Biological Test Items

ZUSAMMENFASSUNG

Die sichere Verwendung von Fachsprache und die Kenntnis von Fachbegriffen sind wesentliche Merkmale für die Expertise bzw. Kompetenz in einer Domäne. Sie sind daher auch für die Modellierung und Erfassung von Kompetenzen von Bedeutung. Da die Kompetenzmodellierung im Allgemeinen auf einer Operationalisierung durch Aufgaben beruht, stellt sich die Frage, inwieweit Fachsprache bzw. Fachbegriffe die Schwierigkeit von Testaufgaben beeinflussen. In dieser Studie wurden Testaufgaben zu drei exemplarischen Konzepten der Biologie hinsichtlich der Verwendung von Fach- und Alltagsbegriffen variiert und mit Schülerinnen und Schülern der Klasse 5 bis 10 getestet. Im Vergleich erweisen sich Testaufgaben mit Fachbegriffen über alle untersuchten Konzepte hinweg als signifikant schwerer als solche mit Alltagsbegriffen. Allerdings unterscheidet sich der schwierigkeitserzeugende Einfluss der Fachbegriffe bei den verschiedenen Konzepten deutlich in seiner Stärke.

Schlüsselwörter: Kompetenzmodellierung, Fachsprache, Fachbegriffe, Testaufgaben, schwierigkeitserzeugende Aufgabenmerkmale

ABSTRACT

The application of terminology and knowledge of technical terms are fundamental attributes for the expertise respectively competency in a domain. Thus they are also relevant for modelling and characterizing competencies. The competency modelling is typically based on test items. Therefore, the question comes up in how far terminology respectively technical terms influence the difficulty of items in achievement tests. In this study items on three exemplary biological concepts have been developed. The items have been varied regarding the use of technical terms and everyday speech. Students of grades 5 to 10 took part in the achievement test. The results show that items with technical terms are significantly more difficult on all reviewed concepts than items containing everyday language. However there is a clear difference within the degree of difficulty caused by the employment of technical terms in all three concepts.

Keywords: competency modelling, terminology, technical terms, test items, item difficulty

1 Einleitung

Die Förderung der Fähigkeit, Fachsprache korrekt und situationsangemessen zu verwenden, ist ein zentrales Anliegen des Fachunterrichts. Dies zeigt sich unter anderem in der Berücksichtigung des Kompetenzbereichs Kommunikation in den zentralen Vorgaben für den Biologieunterricht in den nationalen Bildungsstandards (KMK 2005, 7) und den Einheitlichen Prüfungsanforderungen für die Abiturprüfung (EPA; KMK 2004, 8). Daher sind die im Unterricht und in Lernsituationen stattfindenden Kommunikationsprozesse auch Gegenstand verschiedener Forschungsprojekte in den Naturwissenschaftsdidaktiken (u. a. Kramer 2009; Kulgemeyer & Schecker 2009; Nitz, Nerdel & Prechtel 2010; Burmeister, Schmiemann & Sandmann 2010; Knobloch, Sumfleth & Walpuski 2009), die hier nicht eingehender dargestellt werden sollen. Diese Studie befasst sich mit den Fachbegriffen als ein konstituierendes Element der Fachsprache. Im Mittelpunkt steht dabei ihre Bedeutung als schwierigkeiterzeugendes Aufgabenmerkmal sowie für die Modellierung und Erfassung von Kompetenzen.

2 Theoretischer Hintergrund

2.1 Fachsprache und Fachbegriffe

Die Kommunikation in einer Domäne ist geprägt durch ihre jeweilige Fachsprache. Sie stellt eine mehr oder weniger explizite Übereinkunft der Experten in der jeweiligen Disziplin dar und dient einem

effizienten, kontextunabhängigen und intersubjektiven Austausch von Informationen zwischen den Beteiligten (Roelcke 2005; Buhlmann & Fearn 2000, 13; Wichter 1994, 43). Sie vereinfacht so einerseits die Verständigung innerhalb der Fachgemeinschaft. Andererseits trägt sie zur Abgrenzung gegenüber Laien (Bromme & Rambow 2001) sowie Experten anderer Disziplinen bei (Bromme & Bündler 1994). Die Wissenschafts- und die Alltagssprache (auch Umgangs- oder Gemeinsprache genannt) unterscheiden sich in drei zentralen Merkmalen voneinander (Klein 2003): Erstens sind Wissenschaftssprachen durch charakteristische Textsorten gekennzeichnet (Göpferich 1995), in denen typische Satzbaumuster verwendet werden, z. B. das unpersönliche Passiv. Zu diesen Textsorten gehören beispielsweise wissenschaftliche Aufsätze in Zeitschriften. Zweitens verwenden Wissenschaftssprachen spezifische Symbolschätze, z. B. Formeln, Zeichen und Abkürzungen. Drittens verfügen sie über einen eigenen Wortschatz (vgl. auch Wichter 1994, 43; Fluck 1997, 20). Erst diese typische Lexik erlaubt es, die jeweilige Fachsprache zu identifizieren und von anderen zu unterscheiden (Klein 2003). Der Wortschatz beinhaltet die Fachwörter. Ihnen kommt somit eine besondere Relevanz für die Fachsprache zu (Fluck 1997, 35; Buhlmann et al. 2000, 34). Unter einem Wort wird im linguistischen Sinne „die kleinste bedeutungstragende und zugleich frei verwendbare sprachliche Einheit“ (Roelcke 2005, 51) verstanden. Im Bereich der naturwissenschaftlichen Fachdidaktiken wird allerdings bevorzugt die Bezeich-

nung *Fachbegriff* verwendet (z. B. Kattmann 1992; Berck 1986; Vollmer 1980). Damit soll vor allem deutlich gemacht werden, dass hinter den mehr oder weniger austauschbaren Wörtern Bedeutungen, also Begriffe, stehen. Diese Bedeutungen bleiben bei der ständigen Veränderung der Fachsprache, wie z. B. Wechsel oder Veränderungen von einzelnen Wörtern, weitgehend stabil (Klein 2003). Denn „ein Begriff ist [...] nicht der Name dessen, was ein Begriff erfassen soll, sondern das Erfasste selbst“ (Rohracher 1988, 362).

Die Zuordnung konkreter Wörter zur Fach- bzw. zur Alltagssprache ist nicht so trivial, wie sie zunächst erscheinen mag (Göpferich 1995, 23ff.). Dies ist unter anderem auf die Mehrdeutigkeit vieler Wörter (Vollmer 1980, 43f.) und auf die Abhängigkeit ihrer Bedeutung von der jeweiligen Anwendungssituation zurückzuführen (z. B. Stärke als körperliche Kraft oder als Gemisch aus Amylose und Amylopektin; Knock-out beim Boxen oder in der molekulargenetischen Forschung). Ein mögliches Verfahren, das der komplexen Gliederung von Sprache (Göpferich 1995) Rechnung trägt, ist die Festlegung eines so genannten Fachlichkeitsindex (Wichter 1994). Demnach kann Wörtern ein eher hoher oder ein eher niedriger Fachlichkeitsindex zugeschrieben werden. Die Begriffe der Alltagssprache und die Begriffe der Fachsprache bilden gleichsam die Extrema auf der Skala des Fachlichkeitsindex. Ein hoher Fachlichkeitsindex deutet darauf hin, dass ein Wort vornehmlich von Experten verwendet wird (Klein 2003, 30). Dieser hier definierte Zusammenhang zwischen der Kenntnis bzw.

Verwendung von Fachbegriffen und Expertentum findet sich auch in empirischen Untersuchungen wieder. So konnten beispielsweise Kroß & Lind (2001) für das Thema Genetik zeigen, dass das Vorwissen auch den Gebrauch der Fachsprache in den Selbsterklärungen bei der Auseinandersetzung mit Beispielaufgaben beeinflusst. Sie verglichen hierzu das Lernen von Teilnehmerinnen und Teilnehmern des Auswahlwettbewerbs zur Internationalen Biologieolympiade („Biologieexperten“) und Physikolympiade („Biologienovizen“). Personen mit einem geringen Vorwissen nutzten bei den Selbsterklärungen vor allem Alltags- und Fachbegriffe aus den gegebenen Beispieltextritten sowie textfremde Alltagsbegriffe. Personen mit hohem Vorwissen hingegen verwendeten häufiger Fachbegriffe, die nicht im Lernmaterial enthalten waren. Diese Fachbegriffe gehörten entweder zum Schulstoff oder gingen darüber hinaus. Die Kenntnis und Verwendung von Fachbegriffen scheinen demnach auch einen Teil fachlicher Fähigkeiten auszumachen, weshalb es angemessen erscheint, dies auch im Bereich der Kompetenzmodellierung zu berücksichtigen.

2.2 Kompetenzmodellierung, Leistungsanforderungen und Aufgabenschwierigkeiten

Bevor im Folgenden die Bedeutung von Begriffen für die Schwierigkeit von Testaufgaben (und damit auch für die Erfassung von Kompetenzen) beleuchtet wird, werden zunächst kurz der allgemeine

Rahmen zur Kompetenzmodellierung und die damit verbundenen Aspekte der Leistungsanforderungen und Aufgabenschwierigkeiten dargestellt.

Mit der Einführung nationaler Bildungsstandards hat die Erfassung und Modellierung von Kompetenzen im Bereich des naturwissenschaftlichen Unterrichts mit dem Ziel der Bilanzierung, Verbesserung und Weiterentwicklung von Lernprozessen an Bedeutung gewonnen (Klieme et al. 2003; Klieme & Leutner 2006; KMK 2006). Im Zentrum dessen steht die Entwicklung von Kompetenzmodellen, die systematisch und fachdidaktisch begründet Leistungsanforderungen für die jeweilige Domäne beschreiben (Klieme et al. 2003; Klieme & Steinert 2004; Schecker & Parchmann 2006; Klieme, Maag-Merki & Hartig 2007; Klieme et al. 2006). Kompetenzen werden dabei in Anlehnung an die Definition von Weinert (2002), die sich aus der Expertiseforschung ableitet (vgl. Weinert 2001; Klieme et al. 2003, 72), verstanden „als kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen“ (Klieme et al. 2006, 879). Für die Modellierung und Diagnostik von Kompetenzen ist eine systematische Beschreibung dieser Situationen und Anforderungen von besonderer Relevanz (Hartig & Jude 2007). Allgemein erfolgt dies in Form von Testaufgaben (Helmke & Hosenfeld 2004). Die Testaufgaben bilden die Situationseigenschaften in Form von schwierigkeitsbestimmenden Aufgabenmerkmalen ab (Anforderungsprofil; Klieme et al. 2003). Letztere resultieren aus solchen Eigenschaften von

Testaufgaben, die unterschiedlich hohe Anforderungen an die Testperson stellen und sich in dementsprechend unterschiedlichen Lösungswahrscheinlichkeiten der Aufgaben ausdrücken. In der Zusammenführung von theoretisch begründeten, schwierigkeits erzeugenden Aufgabenmerkmalen und empirischen Lösungswahrscheinlichkeiten von Aufgaben ist es möglich, die zu Grunde liegenden Kompetenzen empirisch fundiert zu beschreiben. In den großen Schulleistungsstudien TIMSS und PISA wurden beispielsweise Aufgaben anhand ihrer empirischen Schwierigkeit unter Berücksichtigung konzeptioneller Überlegungen den Kompetenzniveaus post hoc zugeordnet und so die relevanten Kompetenzen beschrieben (Baumert, Bos & Lehmann 2000; Klieme, Baumert, Köller & Bos 2000; Baumert, Stanat & Demmrich 2001; Adams & Wu 2002). Alternativ zu diesem induktiven Verfahren wurden in verschiedenen Studien auch andere Verfahren angewandt (Helmke et al. 2004). Besonders empfehlenswert erscheint aus Sicht der psychologischen Diagnostik eine deduktive Testkonstruktion (Hartig et al. 2007). Dabei werden a priori theoriegeleitete Annahmen über zu erwartende Aufgabenschwierigkeiten getroffen und anschließend empirisch geprüft. Nach diesem Ansatz wurden beispielsweise bei der Studie „Deutsch Englisch Schülerleistungen International“ (DESI) die Kompetenzniveaus definiert (Hartig 2007). Dieses Vorgehen bietet unter anderem den Vorteil, dass die Ergebnisse wegen der genaueren Kenntnis der zu Grunde liegenden Einflussfaktoren leichter zu generalisieren sind, denn „wer-

den Annahmen über die schwierigkeitsbestimmenden Merkmale einer Aufgabe vor der Erhebung empirischer Daten formuliert, gewinnen diese den Status empirisch prüfbarer Hypothesen“ (Hartig 2007, 89). Dies setzt allerdings voraus, dass eben solche Faktoren oder Merkmale der Aufgaben, die Schwierigkeit zu induzieren vermögen, bereits zuvor bekannt sind. Prenzel, Häußler, Rost & Senkbeil (2002) identifizierten diesbezüglich drei wesentliche Gruppen von Merkmalen: (1.) formale Aufgabenmerkmale, (2.) kognitive Anforderungen beim Lösen der Aufgaben als Prozessvariablen und (3.) Merkmale der für das Lösen der Aufgaben erforderlichen Wissensbasis als Konzeptvariable. Zu den formalen Aufgabenmerkmalen zählen vor allem „äußere“ Merkmale der Aufgabenstellung und der geforderten Lösung. Dazu gehören z. B. das objektiv gut einschätzbare Antwortformat, das nachweislich die Schwierigkeit von Testaufgaben beeinflussen kann, auch wenn die empirischen Befunde teilweise uneinheitlich sind (vgl. Bennett & Ward 1993), oder auch die verwendeten Repräsentationsformen. Die kognitiven Anforderungen beschreiben die bei der Bearbeitung der Aufgabe stattfindenden kognitiven Prozesse, wie beispielsweise das Verarbeiten von Textinformationen, das Ausrechnen und das Aufbauen räumlicher Modelle. Eine wesentliche Rolle spielt außerdem die erforderliche Wissensbasis, also jene Merkmale des Wissens, die für ein erfolgreiches Bewältigen der Aufgabe von Nöten sind. Hierzu zählen u. a. das Faktenwissen und das Terminologische Wissen, als Kenntnis von Begriffen, die für die Aufga-

benbewältigung vorausgesetzt wird. Dieser letztgenannte Aspekt ist im Hinblick auf den Untersuchungsgegenstand dieser Studie von besonderem Interesse. Daher werden im Folgenden drei Studien (in chronologischer Reihenfolge) vorgestellt, bei denen die durch Begriffe verursachte Aufgabenschwierigkeit näher untersucht wurde. Bei den Studien handelt es sich ausnahmslos um post-hoc Reanalysen von Aufgaben aus großen Schulleistungstudien.

Aufgrund der Diskrepanz in den Leistungen schweizerischer Schülerinnen und Schüler in Mathematik und den Naturwissenschaften zwischen der TIMS-Studie (Third International Mathematics and Science Study) und der IEA-Studie (International Association for the Evaluation of Educational Achievement; Moser, Ramseier, Keller & Huber 1997, 46 ff.) wurden die schweizerischen TIMSS-Daten einer eingehenden Analyse unterzogen (Ramseier 1997). Es wurde vermutet, dass die Abweichung in den Naturwissenschaftsleistungen bei der TIMS-Studie unter anderem auf die „begriffliche Schwierigkeit“ zurückzuführen sei, also die stärkere Voraussetzung von Begriffen und Begriffsbezeichnungen in den TIMSS-Aufgaben (Ramseier 1997, 22). Zur Aufklärung dieser Vermutung wurden Experten gebeten, die dort eingesetzten Testaufgaben nachträglich auf einer fünfstufigen Skala bezüglich der begrifflichen Schwierigkeit bezogen auf die Testpopulation einzuschätzen. Die Skala reichte dabei von „einfache, allgemeinverständliche Begriffe“ über „allgemein bekannte Fachbegriffe“ bis zu „spezifische schwierige Fachbe-

griffe als Fremdworte“ (Ramseier 1997, 35, 62). Diese Experteneinschätzungen zeigten eine hohe Übereinstimmung in Faktorenanalyse und Reliabilitätsberechnungen. Auf Grundlage dieser Experteneinschätzungen wurden die beurteilten Aufgaben in drei etwa gleichgroße Gruppen eingeteilt, die mit niedriger (49 Aufgaben), mittlerer (52 Aufgaben) und hoher begrifflicher Schwierigkeit (45 Aufgaben) bezeichnet wurden. Der varianzanalytische Vergleich der relativen Aufgabenschwierigkeit für die Schweiz (d.h. der Differenz zwischen schweizerischer und internationaler Schwierigkeit) zeigte eine hypothesenkonforme Zunahme der Aufgabenschwierigkeit in den drei Gruppen ($F(2,143) = 8,001; p < 0,001; \eta = 0,32$). In einer multiplen Regression konnte diese relative Schwierigkeit zu 25 % auf die begriffliche Schwierigkeit ($\beta = 0,21$) sowie die kognitiven Anforderungen ($\beta = -0,36$) und die internationale Schwierigkeit ($\beta = 0,24$) zurückgeführt werden (Ramseier 1997, 38). Dies galt allerdings nicht für die absolute Schwierigkeit der Aufgaben in der Schweiz. Hierbei spielte vor allem die thematische Schwierigkeit ($\beta = 0,33$) eine Rolle, nicht aber die begriffliche (Ramseier 1997, 41). Gleiches galt auch für die internationale Schwierigkeit der untersuchten TIMSS-Aufgaben. Ramseier (1997) interpretierte diese Befunde dahingehend, dass für schweizerische Schülerinnen und Schüler im Vergleich zu Teilnehmern anderer Länder mit schwierigeren Begriffen (gemäß des Expertenratings) auch eine erhöhte Schwierigkeit einhergehe. Dies erklärte er mit einer geringen Fokussierung des naturwissenschaftlichen Unterrichts

in der Schweiz auf Begriffe und Fachterminologie.

Auch die TIMSS-Testaufgaben zur Oberstufenmathematik und -physik wurden nachträglich auf Basis einer Schwierigkeitseinschätzung durch Experten analysiert (Klieme 2000, 58). Dabei wurde u. a. das sogenannte „qualitative Begriffsverständnis“ als Anforderungsmerkmal der Aufgaben identifiziert. Es umfasst das Verständnis dafür „wie ein Fachbegriff mit anderen Begriffen zusammenhängt, wann er anwendbar ist [...] und was ihn von verwandten Alltagsbegriffen unterscheidet“ (Klieme 2000, 73). Die Befunde zeigen eine hohe Korrelation ($r = 0,54$) mit der Aufgabenschwierigkeit für die Mathematik-, aber keine bei den Physikaufgaben ($r = 0,08; n. s.$). Demnach sind „Mathematikaufgaben aus TIMSS [...] schwierig, wenn sie [u. a.] Verständnis und Kenntnis mathematischer Begriffe [...] voraussetzen“ (Klieme 2000, 82), nicht aber Physikaufgaben.

Im Rahmen der PISA-2000-Studie wurde die Erhebung von Schülerkompetenzen in den Naturwissenschaften durch eine nationale Ergänzungsstudie erweitert, um differenziertere Ergebnisse über die Leistungsfähigkeit deutscher Schülerinnen und Schüler zu erhalten (Neubrand 2004). Die dabei verwendeten Aufgaben wurden in einer Reanalyse mit den internationalen Aufgaben verglichen und auf potenziell schwierigkeits erzeugende Aufgabenmerkmale hin untersucht. Dabei war die Frage leitend, inwieweit die empirischen Itemschwierigkeiten durch bestimmte Aufgabenmerkmale erklärt bzw. vorhergesagt werden können (Prenzel et

al. 2002). Zu diesem Zweck wurden von der nationalen Expertengruppe Kategorien entwickelt, die in den drei bereits erwähnten Gruppen (formale Aufgabenmerkmale, kognitive Anforderungen, erforderliche Wissensbasis) zusammengefasst wurden. Auf Basis der einzelnen Kategorien wurden alle nationalen und internationalen Aufgaben signiert, d. h. geprüft, ob das jeweilige Merkmal darin vorkommt oder nicht. Auf Grundlage dieser Zuordnung konnten Prenzel et al. (2002) teils erhebliche Unterschiede zwischen den Merkmalen der nationalen und internationalen Aufgaben beschreiben. Zur Beantwortung der Forschungsfrage wurden die so zugeordneten Merkmale als Prädiktoren für die empirische Itemschwierigkeit in eine Regressionsanalyse einbezogen. Einige Aufgabenmerkmale wurden allerdings ausgeschlossen, da sie für das Schwierigkeitsmodell nicht geeignet erschienen. Prenzel et al. (2002) begründen dies u. a. mit dem gleichzeitigen gehäuften Auftreten einiger Merkmale, was auf eine Konfundierung hindeute, sowie der inhaltlichen Ähnlichkeit mancher Kategorien. Mit den verbliebenen 15 Aufgabenmerkmalen ließen sich rund 45% der Aufgabenschwierigkeit erklären. Zu den Merkmalen mit den höchsten Regressionsgewichten, und damit auch dem stärksten schwierigkeits erzeugenden Einfluss, gehörten: Etwas ausrechnen ($B = 1,39$; kognitive Anforderungen), Terminologisches Wissen ($B = 1,30$; Wissensbasis) und Freie lange Antwort ($B = 0,85$; formale Aufgabenmerkmale). Demnach scheinen „Aufgaben, in denen das Verständnis oder die Nennung eines bestimmten na-

turwissenschaftlichen Begriffs vorausgesetzt wird, besondere Schwierigkeiten zu bereiten“ (Prenzel et al. 2002, 132). Einschränkung geben sie aber zu bedenken, dass einzelne Aufgaben nur erfolgreich gelöst werden könnten, wenn bestimmte Fähigkeiten *und* die entsprechende Wissensbasis vorhanden seien. Sie schränken daher die inhaltliche Generalisierbarkeit der Befunde, v. a. hinsichtlich der Stärke des schwierigkeits erzeugenden Einflusses einzelner Faktoren, stark ein, betonen aber zugleich die große Bedeutung für zukünftige Testkonstruktionen.

Die dargestellten Befunde deuten klar auf einen Einfluss von Begriffen auf die Schwierigkeit von Testaufgaben hin. Allerdings beschränken sich die bisherigen Untersuchungen auf die Reanalyse von bestehenden Testaufgaben. Angesichts dieser Tatsachen erscheint eine systematische Untersuchung des Einflusses von Fachbegriffen auf die Schwierigkeit von Testaufgaben sinnvoll, bei der die Aufgaben a priori in diesem Merkmal variiert werden.

3 Ziel und Forschungsfragen

Ziel der Studie ist es, den Einfluss von biologischen Fachbegriffen (im Gegensatz zu Alltagsbegriffen) auf die Schwierigkeit von Testaufgaben näher zu charakterisieren. Mit diesem Ziel lassen sich zwei empirisch überprüfbare und aufeinander aufbauende Forschungsfragen formulieren:

1. Unterscheiden sich Testaufgaben mit biologischen Fachbegriffen von sol-

chen mit Alltagsbegriffen in ihrer empirischen Schwierigkeit?

2. Wie groß ist der Einfluss von biologischen Fachbegriffen auf die Schwierigkeit von Testaufgaben bei verschiedenen Konzepten?

4 Methodisches Vorgehen

4.1 Aufgabenentwicklung

Die Testaufgaben wurden im Rahmen der Operationalisierung und empirischen Überprüfung eines Kompetenzstrukturmodells zum biologischen Fachwissen entwickelt. Im Folgenden werden einzelne Merkmale der Modellstruktur soweit kurz dargestellt, wie sie für die Aufgabenentwicklung unter den hier fokussierten Forschungsfragen von Bedeutung sind. Weitere Details sind an anderer Stelle (Schmiemann 2010) ausführlicher dargestellt.

Da sich domänenspezifische Kompetenzen immer auf konkrete Inhalte beziehen (Klieme et al. 2003, 75) und sich diese in den Naturwissenschaften vor allem als fachliche Konzepte manifestieren (KMK 2005, 7), wurden Testaufgaben entwickelt, die konzeptuelles biologisches Wissen erfassen. Der Kategorisierung von Prenzel et al. (2002) folgend, ist das konzeptuelle Wissen der Kategorie der erforderlichen Wissensbasis zuzuordnen. Für diese Studie wurde diese Konzeptkategorie zusätzlich so differenziert, dass unterschiedliche Niveaus erfasst werden können. Im Gegensatz zu den übergeordneten Merkmalen von Prenzel et al. (2002) beziehen sich

die hier berücksichtigten Niveaus konkret auf die für das Kompetenzmodell exemplarisch ausgewählten biologischen Konzepte Blutkreislauf, Entwicklung und Vererbung (s. u.). Bei der Auswahl dieser Konzepte waren insbesondere zwei Überlegungen leitend: Erstens werden die drei Konzepte i. A. sowohl im Biologieunterricht unterschiedlicher Schulformen als auch in verschiedenen Jahrgangsstufen behandelt. Mit diesem Auswahlkriterium verband sich die Erwartung, mittels einer breit gestreuten Stichprobe ein umfangreiches Leistungsspektrum erfassen zu können, ohne einzelne Teilpopulationen systematisch zu benachteiligen. Zweitens sind für diese Konzepte typische Verständnisschwierigkeiten und Schülerfehler bekannt (u. a. Arnaudin & Mintzes 1986; Shepardson 1997; Lewis 2000). Ihre Berücksichtigung trägt der Forderung von Klieme et al. (2003) Rechnung, „konkrete Ausformulierungen und Operationalisierungen des Kompetenzbegriffs [...] auf dem Theorie- und Erkenntnisstand der Fachdidaktiken aufzubauen“ (S. 75). Diese bekannten Verständnisschwierigkeiten und Schülerfehler bildeten den Ausgangspunkt für die Festlegung von Niveaus des konzeptuellen Wissens für jedes der drei Konzepte. Für das Konzept Blutkreislauf wurden drei Niveaus (Durchblutung, Blutkreislauf, Körper- und Lungenkreislauf), für das Konzept Entwicklung ebenfalls drei Niveaus (Entwicklungsstadium, Entwicklungsschritt, Entwicklungszyklus) und für das Konzept Vererbung fünf Niveaus (Verwandtschaftsähnlichkeit, Erworbene Eigenschaften, Erbmaterial, Variabilität, Vererbungsregeln) mit

jeweils ansteigender Schwierigkeit angenommen (Schmiemann 2010). Um diese Annahmen empirisch prüfen zu können, wurden zur Operationalisierung jedes postulierten Niveaus mehrere Testaufgaben konstruiert. Diese Entwicklung erfolgte in einem mehrstufigen Verfahren von zunächst offenen Aufgaben über halboffene Aufgaben zu geschlossenen Mehrfachwahlaufgaben mit jeweils einem Attraktor und drei Distraktoren (vgl. Tabelle 1). Sie enthielten jeweils einen Aufgabenstamm mit Frage und die vier Antwortmöglichkeiten. Grafiken oder andere Bildinformationen wurden nicht gegeben, so dass die Aufgaben hinsichtlich formaler Merkmale als weitgehend konstant betrachtet werden können.

Um die hier im Fokus stehenden Forschungsfragen beantworten zu können, wurden die Testaufgaben außerdem so konstruiert, dass sie entweder biologische Fachbegriffe oder Alltagsbegriffe enthielten. Diese Integration konnte sich sowohl auf den Aufgabenstamm als auch auf die Antwortalternativen beziehen (vgl. Tabelle 1). Aufgrund der bekannten Schwierigkeit einer eindeutigen Zuordnung bei mehrstufigen Einschätzungen (s. o. bzw. Ramseier 1997) wurde hier eine dichotome Differenzierung in die Extrema gewählt, in der Erwartung konsistentere Befunde zu erhalten. Der Kategorie „Fachbegriffe“ wurden solche Begriffe zugeordnet, die in den typischen Biologie-schulbüchern der Unter- und Mittelstufe

Tab. 1: Beispiele für Testaufgaben. Unterschiede zwischen den Aufgaben durch Alltags- bzw. Fachbegriffe sind hervorgehoben

Testaufgabe mit Alltagsbegriffen	Testaufgabe mit Fachbegriffen
<p>v13 Ein Gärtner hat Blumen mit weißen Blüten. Er färbt die Blüten blau und <i>vermehrt</i> diese Pflanzen. Welche Farbe haben die Blüten der <i>entstehenden Pflanzen</i> am wahrscheinlichsten?</p> <p><input type="checkbox"/> Die meisten Blüten sind weiß. <input type="checkbox"/> Es sind in etwa gleich viele Blüten blau und weiß. <input type="checkbox"/> Die meisten Blüten sind hellblau. <input type="checkbox"/> Die meisten Blüten sind blau.</p>	<p>v14 Ein Gärtner hat Blumen mit weißen Blüten. Er färbt die Blüten blau und <i>kreuzt</i> diese Pflanzen. Welche Farbe haben die Blüten der <i>Filialgeneration</i> am wahrscheinlichsten?</p> <p><input type="checkbox"/> Die meisten Blüten sind weiß. <input type="checkbox"/> Es sind in etwa gleich viele Blüten blau und weiß. <input type="checkbox"/> Die meisten Blüten sind hellblau. <input type="checkbox"/> Die meisten Blüten sind blau.</p>
<p>e4 Welche Aussage beschreibt einen Abschnitt im Leben eines Schmetterlings richtig?</p> <p><input type="checkbox"/> In den Schmetterlingseiern findet eine vollständige <i>Verwandlung</i> statt. <input type="checkbox"/> Die Schmetterlings<i>raupen</i> legen Eier. <input type="checkbox"/> <i>Der Schmetterling</i> durchläuft eine <i>Verwandlung</i> zur <i>Raupe</i>. <input type="checkbox"/> <i>Der Schmetterling</i> legt Eier.</p>	<p>e5 Welche Aussage beschreibt einen Abschnitt im Leben eines Schmetterlings richtig?</p> <p><input type="checkbox"/> In den Schmetterlingseiern findet eine vollständige <i>Metamorphose</i> statt. <input type="checkbox"/> Die Schmetterlings<i>larven</i> legen Eier. <input type="checkbox"/> <i>Die Schmetterlingsimago</i> durchläuft eine <i>Metamorphose</i> zur <i>Larve</i>. <input type="checkbox"/> <i>Die Schmetterlingsimago</i> legt Eier.</p>

vorkommen und üblicherweise nicht in der Alltagssprache verwendet werden (z. B. Filialgeneration, Froschlarve). Begriffe die sowohl in der Alltags- als auch in der Fachsprache genutzt werden, wurden bei gleicher oder ähnlicher Bedeutung der Kategorie „Alltagsbegriffe“ zugeordnet (z. B. Herz, Kaulquappe). Begriffe mit unterschiedlichen Bedeutungen in Alltags- und Fachsprache wurden durch Synonyme ersetzt, um eine Konfundierung auszuschließen. Bewusst ausgeschlossen wurde das Fachvokabular der universitären Lehrwerke und wissenschaftlichen Publikationen mit seinen vor allem lateinischen und griechischen Wortstämmen, da die Testaufgaben auf den Biologieunterricht der Sekundarstufe I fokussieren. Ausnahmen ergeben sich nur insoweit, als einige dieser Begriffe, wie beispielsweise „Metamorphose“, auch in Mittelstufenschulbüchern verwendet werden. In diesem Sinne handelt es sich genau genommen bei den hier untersuchten „Fachbegriffen“ um die Fachbegriffe der Unterrichtssprache. Die mit Fachbegriffen formulierten Testaufgaben setzten also i. S. der terminologischen Schwierigkeit bei Prenzel et al. (2002) die Kenntnis der entsprechenden Begriffe für eine erfolgreiche Bearbeitung voraus. Die Zuordnung der Begriffe bzw. Aufgaben zu den Kategorien wurde von einer Expertengruppe aus Biologiedidaktikern und erfahrenen Biologielehrkräften diskutiert, bis eine einstimmige Einteilung vorgenommen werden konnte. Dabei wurde versucht, inhaltlich analoge Aufgabenpaare zu erhalten, von denen die eine Aufgabe mit Fachbegriffen, die an-

dere mit Alltagsbegriffen formuliert ist (vgl. Tabelle 1). Zum Teil war dies allerdings nicht möglich, da für einige Begriffe, die für die Operationalisierung des konzeptuellen Wissens unerlässlich sind (z. B. Venen, Arterien beim Konzept Blutkreislauf), keine analogen Alltagsbegriffe bestimmt werden konnten. In solchen Fällen wurde versucht, gleichwertige, d. h. auf dem gleichen Niveau des konzeptuellen Wissens befindliche, Aufgaben mit Alltagsbegriffen zu konstruieren. In einigen wenigen Fällen war auch dies aus inhaltlichen Gründen, die im Rahmen der Diskussion näher erläutert werden, nicht möglich, so dass sich insgesamt eine ungerade Anzahl an Testaufgaben ergibt.

4.2 Design und Stichprobe

Die insgesamt 129 Testaufgaben wurden in einem Multimatrix-Design (Walter 2005) auf acht Testhefte verteilt, um die Bearbeitung innerhalb einer Unterrichtsstunde zu ermöglichen. Jedes Testheft enthielt 32 bzw. 33 Testaufgaben, wobei jede Aufgabe in zwei Testheften vorkam. Um eine Stabilität der Anker auch beim möglichen Ausfall einzelner Aufgaben zu gewährleisten, waren die Aufgaben so verteilt, dass jedes Heft mit vier weiteren Heften verbunden ist. Die so verteilten Aufgaben wurden von über dreitausend ($N = 3337$) Schülerinnen (50 %) und Schülern der Jahrgangsstufen fünf bis zehn aus Haupt-, Real- und Gesamtschulen sowie Gymnasien innerhalb einer Unterrichtsstunde unter Aufsicht der Lehrkraft bearbeitet.

4.3 Bestimmung der Aufgabenschwierigkeit

Bedingt durch das Multimatrix-Design wurden nicht alle Testaufgaben von allen Probanden bearbeitet, so dass der Anteil richtiger Antworten nicht direkt als Lösungswahrscheinlichkeit im Sinne der klassischen Testtheorie interpretiert werden kann. Die Bestimmung der Aufgabenschwierigkeiten und Personenfähigkeiten erfolgte daher mit dem probabilistischen Rasch-Modell (Rasch 1960; Überblick z. B. bei Rost 2004). Aufgrund des dichotomen Antwortmusters (korrekt beantwortet vs. falsch beantwortet) wurde das 1-Parameter-Modell (1PL) verwendet (Embretson & Reise 2000). Die Aufgabenschwierigkeiten (Itemparameter) wurden dabei mit der Software Conquest von ACER nach der marginalen Maximum-Likelihood-Methode bestimmt (Adams & Wu 2007), da sie auch bei umfangreicheren designbedingten Datenlücken eine zuverlässige Parameterschätzung gestattet (Rost 2004). Ein weiterer Vorteil dieses Testmodells liegt darin, dass die Personenfähigkeit und die Aufgabenschwierigkeit auf einer gemeinsamen Logit-Skala liegen und daher direkt mit einander verglichen werden können (Moosbrugger 2007).

4.4 Prüfung der Forschungsfragen

Um zu überprüfen, ob sich die Schwierigkeit von Testaufgaben mit Fachbegriffen von solchen mit Alltagsbegriffen unterscheidet (Forschungsfrage 1), wurden die ermittelten Aufgabenschwie-

rigkeiten in einem T-Test mit einander verglichen.

Zur Quantifizierung des Einflusses von Fachbegriffen auf die Aufgabenschwierigkeit (Forschungsfrage 2) wurden Regressionen berechnet. Hierzu wurden für das Vorkommen von Alltags- bzw. Fachbegriffen in den Testaufgaben Dummy-Variablen angelegt, die als erklärende Variable in das Regressionsmodell eingehen (vgl. Prenzel et al. 2002). Der Determinationskoeffizient R^2 kann dann als Maß dafür interpretiert werden, welcher Anteil der Aufgabenschwierigkeit auf die Verwendung der Fachbegriffe zurückgeführt werden kann.

5 Ergebnisse

5.1 Modell- und Aufgabenprüfung

Um die Gültigkeit des verwendeten Rasch-Modells bezogen auf die empirischen Antwortmuster zu überprüfen, werden verschiedene Indikatoren herangezogen. Dabei nehmen die gewichteten Abweichungsquadrate zwischen den erwarteten und den beobachteten Häufigkeiten ($wMNSQ$) eine zentrale Stellung ein. Die ermittelten Infit-Werte ($0,86 \leq wMNSQ \leq 1,11$) liegen innerhalb der typischen Grenzwerte (Bond & Fox 2001, 179; Adams et al. 2002, 105). Die WLE-Reliabilität liegt mit einem Wert von 0,73 im guten Bereich.

Insgesamt weisen die Testaufgaben eine für die Stichprobe angemessene Schwierigkeit auf. Dies zeigt sich sowohl an der mittleren Personenfähigkeit von $M_0 = +0,15$ als auch an der Verteilung der Item- und Per-

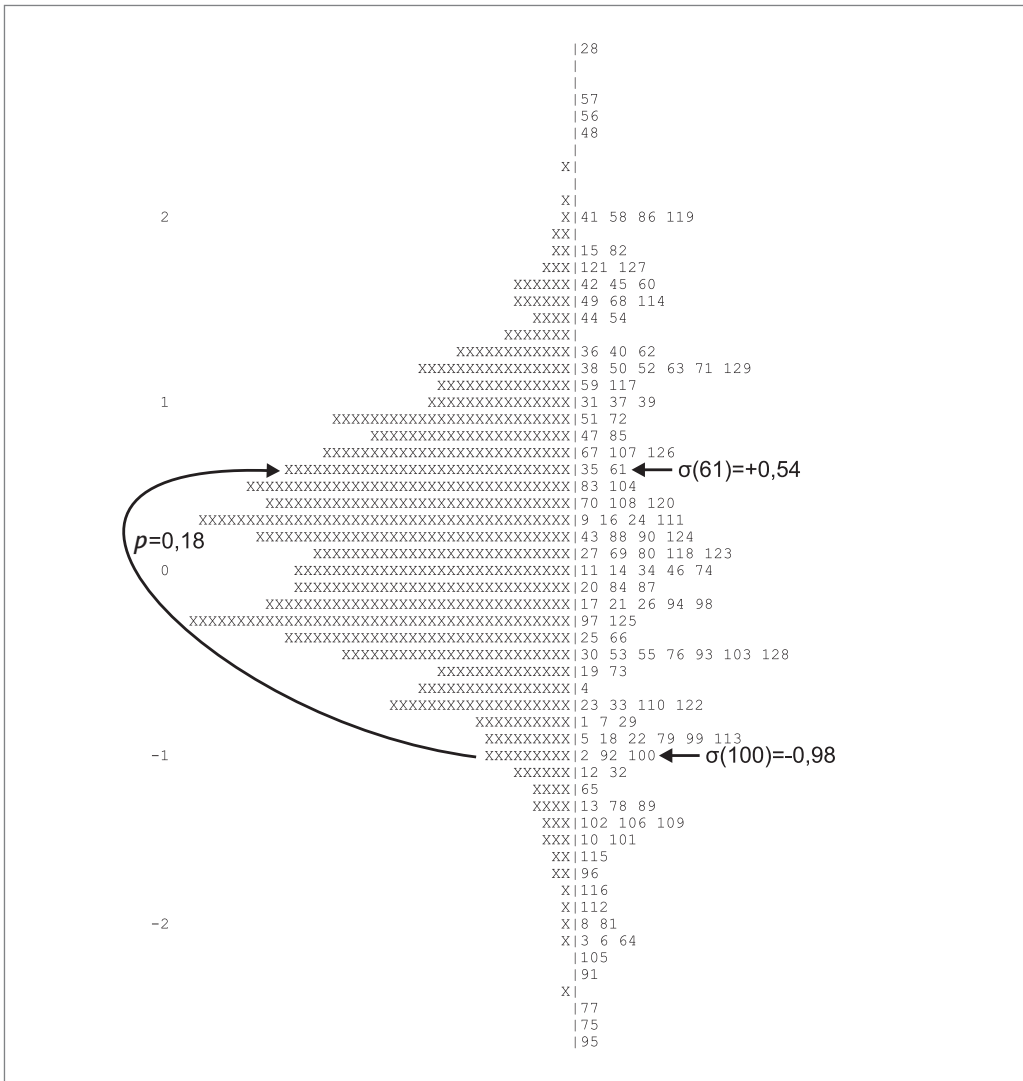


Abb. 1: Verteilung der Item- und Personenparameter. Erläuterungen im Text.

sonenparameter in der Wright-Map (Abbildung 1). Diese zeigt auf der linken Seite die Verteilung der Personenparameter, wobei jedes Kreuz 5,2 Fälle repräsentiert. Auf der rechten Seite ist die Verteilung der Aufgaben dargestellt. Die Zahlen sind die internen Nummern der Aufgaben. Je schwieriger eine Aufgabe ist, desto weiter oben wird sie in der Wright-Map abgetra-

gen. Leichte Aufgaben werden hingegen weiter unten dargestellt. Aufgrund der gemeinsamen Logit-Skala (kleine Zahlen ganz links von -2 bis +2; vgl. Moosbrugger 2007) gilt dies in analoger Weise auch für die Personenfähigkeiten. Personen mit hohen Fähigkeiten sind oben, Personen mit geringen Fähigkeiten unten dargestellt. Eine Person löst eine Aufgabe mit einer

Wahrscheinlichkeit von 50 %, wenn ihr Personenparameter dem Itemparameter entspricht. Sie sind dann in der Wright-Map auf gleicher Höhe dargestellt.

Eine Person mit einem Fähigkeitswert von $\theta = -1,00$ löst die Aufgabe mit Nummer 100 ($\sigma = -0,98$; s. Pfeil) folglich mit einer Wahrscheinlichkeit von etwa 50 %. Die Wahrscheinlichkeit, dass dieselbe Person die schwierigere Aufgabe Nummer 63 ($\sigma = +0,54$; s. Pfeil) löst, liegt nur bei 18 %. Aufgrund der statistischen Kennwerte und der Schwierigkeitsverteilung der Aufgaben ergibt sich keine Notwendigkeit, Aufgaben zu verwerfen, so dass alle Aufgaben in der Auswertung berücksichtigt werden können.

Um außerdem auszuschließen, dass die Schwierigkeit der Aufgaben durch nicht systematisch variierte Aufgabenmerkmale beeinflusst wird, wurden auch Korrelationen bzw. Mittelwertunterschiede für solche möglichen Faktoren berechnet. Dabei zeigt sich kein statistisch bedeutsamer Zusammenhang zur Textlänge des Aufgabenstamms ($r = 0,11$; $p = 0,24$), der Antworten ($r = -0,03$; $p = 0,73$) und der Gesamtaufgabe ($r = 0,003$; $p = 0,73$). Auch konnte weder bezüglich der Position der richtigen Antwort bei jeder einzelnen Aufgabe ($F(3,125) = 0,66$; $p = 0,58$) noch der Position der Aufgabe im Testheft ($F(1,127) = 0,04$; $p = 0,85$) ein Zusammenhang zur Aufgabenschwierigkeit nachgewiesen werden. Zwischen den 16 Aufgabenblöcken, die für das Matrix-Design gebildet wurden, besteht ebenfalls kein Schwierigkeitsunterschied ($F(15,113) = 0,24$; $p = 1,00$). Die Testaufgaben wurden also nachweislich nicht durch

diese Aufgabenmerkmale in ihrer Schwierigkeit beeinflusst. Allerdings wurden sie für die intendierte Kompetenzmodellierung systematisch bezüglich ihres Niveaus des konzeptuellen Wissens verändert, was sich erwartungsgemäß auch in den empirischen Aufgabenschwierigkeiten niederschlägt. So kann ein Anteil von $R^2_{\text{adj}} = 0,40$ der Aufgabenschwierigkeit beim Konzept Blutkreislauf auf die angenommenen drei Niveaus des konzeptuellen Wissens zurückgeführt werden. Bei den Konzepten Entwicklung und Vererbung erwiesen sich nicht alle angenommenen konzeptuellen Niveaus als empirisch haltbar. Sie wurden daher unter Berücksichtigung inhaltlicher Überlegungen teilweise zusammengeführt, so dass sich empirisch differenzierbare Niveaus ergaben. Die verbliebenen zwei Niveaus beim Konzept Entwicklung (Entwicklungsprozess, Entwicklungszyklus) sind für $R^2_{\text{adj}} = 0,48$ der empirischen Aufgabenschwierigkeit verantwortlich. Die drei neuen Niveaus im Konzept Vererbung (Merkmalsweitergabe, Erbanlagen, Vererbungsregeln) erklären $R^2_{\text{adj}} = 0,40$ der empirischen Aufgabenschwierigkeit. Insgesamt wird erwartungsgemäß ein erheblicher Teil der Schwierigkeit der Aufgaben durch die jeweiligen Niveaus des konzeptuellen Wissens bestimmt.

5.2 Schwierigkeitsunterschiede der Begriffe

Im statistischen Mittel liegt die Schwierigkeit von Aufgaben mit Fachbegriffen um 0,80 logits über der Schwierigkeit von Aufgaben, die Alltagsbegriffe enthal-

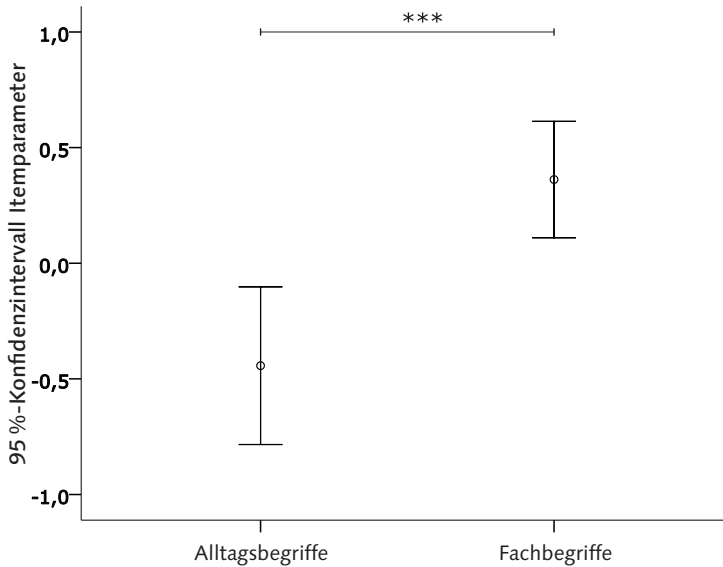


Abb. 2: Mittlere Aufgabenschwierigkeiten (Itemparameter) für Aufgaben mit Alltagsbegriffen und mit Fachbegriffen.

Tab. 2: Unterschiede in den mittleren Aufgabenschwierigkeiten (Itemparameter) für drei Konzepte

Konzept	Mittelwert (Alltagsbegriffe)	Mittelwert (Fachbegriffe)	Mittelwertdifferenz	Signifikanzniveau	Effektstärke
Blutkreislauf	-0,36	+0,33	0,69	$p \leq 0,05$	mittel ($d=0,59$)
Entwicklung	-0,50	+0,47	0,97	$p \leq 0,05$	stark ($d=0,82$)
Vererbung	-0,69	+0,43	1,12	$p \leq 0,001$	stark ($d=0,99$)

ten ($T(127) = -3,874$; $p \leq 0,001$; vgl. Abbildung 2). Diese Differenz entspricht einem Unterschied in der Lösungswahrscheinlichkeit von etwa 20 Prozentpunkten und einem Effekt mittlerer Stärke ($d=0,69$). Dieser Mittelwertunterschied in den Aufgabenschwierigkeiten lässt sich nicht nur für alle Aufgaben insgesamt, sondern auch jeweils getrennt für die Aufgaben der drei

untersuchten Konzepte nachweisen (Tabelle 2).

Bezugnehmend auf die erste Forschungsfrage kann also festgehalten werden, dass sich die empirische Schwierigkeit von Testaufgaben mit Fachbegriffen von solchen mit biologischen Alltagsbegriffen statistisch signifikant unterscheidet. Dabei werden Aufgaben mit Fachbegriffen

im Mittel seltener gelöst als solche mit Alltagsbegriffen.

5.3 Einflusstärke der Fachbegriffe

Die Ergebnisse der Regressionsanalyse zur Quantifizierung des schwierigkeits-erzeugenden Einflusses der Fachbegriffe (Forschungsfrage 2) spiegeln die Mittelwertdifferenzen und die Effektstärken bei den verschiedenen Konzepten wider (Tabelle 3). Dabei zeigen sich deutliche Unterschiede. So ist die Verwendung von Fachbegriffen beim Konzept Vererbung für rund 20 % ($R^2 = 0,196$; $F(1,37) = 8,997$; $p = 0,005$) der Aufgabenschwierigkeit verantwortlich, während es beim Konzept Blutkreislauf nur 8 % ($R^2 = 0,082$; $F(1,61) = 5,465$; $p = 0,023$) sind. Beim Konzept Entwicklung lassen sich etwa 15 % ($R^2 = 0,154$; $F(1,25) = 4,551$; $p = 0,043$) der Aufgabenschwierigkeit auf die Verwendung von Fachbegriffen zurückführen. Der Einfluss der Verwendung von biologischen Fachbegriffen auf die Schwierigkeit von Testaufgaben variiert in Abhängigkeit vom jeweiligen Konzept. Für die hier untersuchten Konzepte und Aufgaben liegt der Einfluss zwischen

acht (Blutkreislauf) und 20 Prozent (Vererbung) der jeweiligen Gesamtschwierigkeit. Damit fällt er in jedem Konzept niedriger aus, als der Einfluss der Niveaus des konzeptuellen Wissens. Die Tatsache, dass die empirische Aufgabenschwierigkeit auf a priori definierte Merkmale der Testaufgaben zurückgeführt werden kann, kann außerdem als Validitätsnachweis für das Testinstrument interpretiert werden (Borsboom, Mellenbergh & van Heerden 2003; Embretson 1998).

6 Diskussion & Ausblick

Es hat sich gezeigt, dass die Verwendung von Fachbegriffen (im Gegensatz zu Alltagsbegriffen) in biologischen Testaufgaben einen Einfluss auf die Aufgabenschwierigkeit hat. Dies konnte für drei unterschiedliche biologische Konzepte (Blutkreislauf, Entwicklung, Vererbung) nachgewiesen werden. Außerdem unterscheidet sich der schwierigkeits-erzeugende Einfluss der Fachbegriffe bei den drei untersuchten Konzepten. Aus diesen Befunden ergeben sich insbesondere Implikationen für die Leistungsmessung und Kompetenzmodellierung.

Tab. 3: Einfluss von Fachbegriffen auf die Aufgabenschwierigkeit als Bestimmungsmaß R^2 bei verschiedenen Konzepten

Konzept	Bestimmungsmaß R^2	Signifikanzniveau
Blutkreislauf	0,082	$p < 0,05$
Entwicklung	0,154	$p < 0,05$
Vererbung	0,196	$p < 0,01$
gesamt	0,106	$p < 0,001$

6.1 Bedeutung für die Leistungsmessung

Die Messung von Leistungen setzt geeignete Testinstrumente voraus. Die gezielte Konstruktion von Testaufgaben ist dabei von großer Bedeutung (vgl. u. a. Bühner 2006, 45ff.; Lienert & Raatz 1994). Die Ergebnisse dieser Studie machen deutlich, dass Testaufgaben weitgehend systematisch a priori hinsichtlich des Vorkommens von Alltags- bzw. Fachbegriffen variiert werden können. Allerdings hat sich bei der Entwicklung der Testaufgaben auch gezeigt, dass es bisweilen schwierig ist, analoge Paare von Alltags- und Fachbegriffen festzulegen (vgl. Göpferich 1995, 23ff.). Diese Problematik resultiert v. a. aus der immanenten Funktion von Fachbegriffen innerhalb der Fachsprache als eindeutige und prägnante Bezeichnungen von mehr oder weniger komplexen Sachverhalten (Fluck 1997, 35; Buhlmann et al. 2000, 34). Diese Funktion ist von wesentlicher Bedeutung für den eingangs erwähnten effizienten, kontextunabhängigen und intersubjektiven Austausch von Informationen mittels Fachsprache (u. a. Roelcke 2005; Buhlmann et al. 2000, 13; Wichter 1994, 43). Sie erklärt aber auch, warum es z. T. nur längere definitorische Umschreibungen von Fachbegriffen in Alltagssprache gibt, aber keine gleichbedeutenden Alltagsbegriffe. Im Hinblick auf zukünftige Untersuchungen, z. B. für andere Inhaltsbereiche, mag es daher sinnvoll erscheinen, nur analoge Aufgabenpaare zu verwenden, die sich ausschließlich in den verwendeten Alltags- bzw. Fachbegriff-

fen unterscheiden. Ein solches Vorgehen würde aufgrund der hoch systematischen Aufgabenvariation scheinbar die Aussagekraft der Befunde weiter erhöhen. Tatsächlich würde aber die Validität dadurch herabgesetzt, da solche Fachbegriffe systematisch ausgeschlossen werden müssten, für die keine gleichbedeutenden Alltagsbegriffe existieren.

Der schwierigkeiterzeugende Einfluss der Fachbegriffe gestattet es überdies, gezielt die Leistungsanforderungen der Testaufgaben mit diesem Aufgabenmerkmal zu beeinflussen. So kann durch die Integration von Fachbegriffen die Schwierigkeit der Aufgaben erhöht und so eine anspruchsvollere Leistung erfasst werden. In dieser Studie wurde die Kenntnis der Fachbegriffe für die erfolgreiche Bewältigung der Aufgabe vorausgesetzt, was der Kategorie des Terminologischen Wissens als Merkmal der erforderlichen Wissensbasis in der Aufgabenreanalyse von Prenzel et al. (2002) entspricht. In weiteren Studien könnten Fachbegriffe auch unter anderen Perspektiven untersucht werden wie z. B. die Anwendung in verschiedenen Kontexten (vgl. Bromme et al. 1994), als definitorisches Wissen oder die Fähigkeit, verschiedene Fachbegriffe zu einander in Beziehung zu setzen (vgl. Klieme 2000). Unabhängig von der genauen Integration der Fachbegriffe in Testaufgaben, erscheint es aus Perspektive der hier gewonnenen Erkenntnisse zum schwierigkeiterzeugenden Einfluss von Fachbegriffen auf Testaufgaben angeraten, bei zukünftigen Testaufgabenentwicklungen die Verwendung von Fachbegriffen in Bezug auf die Aufgabenschwierigkeit zu

beachten. Idealerweise wird eine unsystematische Variation der Fachbegriffe von vorne herein vermieden um unerwünschte Einflüsse auf die Aufgabenschwierigkeit auszuschließen. Alternativ kann zumindest post hoc die Stärke des Einflusses empirisch geprüft und als Korrekturfaktor in statistischen Modellen berücksichtigt werden.

Über den grundsätzlich schwierigkeiterzeugenden Einfluss von Fachbegriffen in Testaufgaben hinaus hat sich gezeigt, dass dieser sich bei verschiedenen Konzepten in seiner Stärke unterscheidet. Insbesondere der im Verhältnis zu den anderen Konzepten geringe Einfluss auf die Aufgabenschwierigkeit beim Konzept Blutkreislauf ist auffällig. Eine mögliche Erklärung hierfür liegt in den entsprechenden Fachbegriffen (z. B. Vene) selbst. Sie werden im Vergleich zu den Fachbegriffen der anderen Konzepte auch gelegentlich in der Alltagssprache verwendet. Dementsprechend verfügen sie im Verhältnis zu den Fachbegriffen der anderen Konzepte über einen insgesamt niedrigeren Fachlichkeitsindex, weshalb sie in den Testaufgaben nicht so deutlich schwierigkeiterzeugend wirken. Im Gegensatz dazu werden die hier berücksichtigten Fachbegriffe im Konzept Vererbung (z. B. rezessiv, homozygot) vor allem zum fachlichen Austausch genutzt, so dass ihnen – im Vergleich – ein höherer Fachlichkeitsindex zuzuordnen ist. Betrachtet man den Fachlichkeitsindex als eine Gesamtskala über mehrere Konzepte hinweg, so liegen die hier bei einzelnen Konzepten als Fachbegriffe festgelegten Begriffe auf unterschiedlichen Niveaus.

6.2 Bedeutung für die Kompetenzmodellierung

Entsprechend der in Abschnitt 2.2 geführten Argumentation zur Bedeutung von Aufgabenmerkmalen für die Kompetenzmodellierung (vgl. auch Hartig et al. 2007) ergeben sich auch diesbezüglich Konsequenzen aus den Studienergebnissen: Der schwierigkeiterzeugende Einfluss von Fachbegriffen kann dahingehend interpretiert werden, dass die Kenntnis von Fachbegriffen im Sinne des Terminologischen Wissens bei Prenzel et al. (2002) einen relevanten Kompetenzfaktor darstellen. Es erscheint daher sinnvoll, die Berücksichtigung einer solchen Dimension in Kompetenzmodellen näher zu prüfen. Das Kompetenzstrukturmodell zum biologischen Fachwissen (Schmiemann 2010), das als Grundlage für die hier dargestellte Aufgabenentwicklung diente, beinhaltet sie unter dem Begriff des „begrifflich-sprachlichen Wissens“ und differenziert in die Komponenten Alltagsbegriffe und Fachbegriffe. Für zukünftige Untersuchungen wären auch noch zusätzliche Graduierungen im Sinne des Fachlichkeitsindex (Wichter 1994) möglich, wobei eine empirische Differenzierung aufgrund der bisherigen Erfahrungen (vgl. z.B. Ramseier 1997) schwierig erscheint. Eine möglicherweise vielversprechende Komponente könnten die hier ausgeklammerten universitären Fachbegriffe als schwierigstes Extremum sein, da so auch das Kompetenzspektrum über die hier fokussierte Sekundarstufe I hinaus erfasst werden könnte.

Auch wenn die Ergebnisse dieser Untersuchung als deutliche Hinweise dafür verstanden werden können, Begriffe als eine (mögliche) Dimension in Kompetenzmodellen zu berücksichtigen, sei doch aufgrund der Erfahrungen in dieser Studie auf eine Problematik hingewiesen: Zumindest bezogen auf den Bereich des Fachwissens besteht zwischen (Fach)Begriffen und Konzepten naturgemäß ein enger Zusammenhang (vgl. auch Überschneidungen der Kategorien bei Prenzel et al. 2002). Aus diesem Grunde hat es sich – vor allem im Hinblick auf die Aufgabenkonstruktion – bisweilen als schwierig erwiesen, diese Dimensionen ganz klar zu trennen. So gehen höhere Niveaus des konzeptuellen Wissens i. A. auch mit entsprechenden Fachbegriffen eines höheren Fachlichkeitsindex einher. Dennoch erscheint, insbesondere aus theoretischer Perspektive, eine möglichst klare Trennung und Charakterisierung dieser Dimensionen erstrebenswert.

Unabhängig von dieser Problematik machen die Ergebnisse dennoch deutlich, welche essentielle Bedeutung Fachbegriffe für das Fachwissen haben, auch wenn erwartungsgemäß ihr empirischer Einfluss anteilig unter dem der jeweiligen Konzepte liegt. Damit kommt den Fachbegriffen auch eine besondere Relevanz für das Lernen zu (vgl. z. B. Buhlmann et al. 2000, 45). Der besonderen Bedeutung von Fachbegriffen für den naturwissenschaftlichen Unterricht wurde seitens der Fachdidaktiken bereits in den 80er-Jahren intensiv Rechnung getragen. Im Fokus standen dabei vor allem die Begriffsbildung und die Isolierung zentraler Begriffe mit dem Ziel der Reduzierung (u. a. Vollmer 1980; Scha-

efer 1980; Cassels & Johnstone 1985; Berck 1986; Graf 1989; Entrich & Staeck 1992) ohne Bezug zu Aufgabenanforderungen oder Kompetenzen. Erfreulicherweise gibt es auch aktuell im Bereich der Naturwissenschaften einige praxisorientierte Ansätze, die versuchen, der Bedeutung von Fachbegriffen Rechnung zu tragen (vgl. Krischer 2007; Meloefski 2007; Wellington & Osborne 2001). Aber auch seitens des Deutschunterrichts wurde die Problematik von Fachbegriffen in Sachtexten anderer Fächer erkannt (u. a. Rudolph 2007). Hier liegt, gerade im Hinblick auf die Problematik des relativ großen Anteils deutscher Schülerinnen und Schülern im unteren Leistungsbereich der Lesekompetenz (vgl. z. B. Baumert et al. 2001), noch ein großes Entwicklungspotenzial für die Fachdidaktiken und Sprachwissenschaften.

Danksagung

Der Autor dankt den anonymen Gutachtern für wertvolle und konstruktive Hinweise zu einer früheren Version dieses Beitrags und dem BMBF für die Unterstützung dieser Forschungsarbeit im Rahmen des Projektes Biologie im Kontext (bik).

Literatur

- Adams, R.J. & Wu, M.L. (2007). The Mixed-Coefficients Multidimensional Logit Model. A Generalized Form of the Rasch Model. In Davier, M. von & Carstensen, C.H. (Hrsg.), *Multivariate and Mixture Distribution Rasch Models. Extensions and Applications* (S. 57–75). New York, NY: Springer.
- Adams, R.J. & Wu, M.L. (Hrsg.) (2002). *PISA 2000 Technical Report*, Paris: OECD.
- Arnaudin, M.W. & Mintzes, J.J. (1986). The Cardiovascular System: Children's Conceptions and Misconceptions. *Science and Children*, 23(5), 48–51.
- Baumert, J., Bos, W. & Lehmann, R. (2000). Untersuchungsgegenstand und Fragestellungen. In Baumert, J., Bos, W. & Lehmann, R. (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit (S. 19–30). Opladen: Leske & Budrich.
- Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000. Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P. et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15–68). Opladen: Leske & Budrich.
- Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P. et al. (Hrsg.) (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*, Opladen: Leske & Budrich.
- Bennett, R.E. & Ward, W.C. (Hrsg.) (1993). *Construction versus choice in cognitive measurement. Issues in constructed response, performance testing, and portfolio assessment*, Hillsdale, NJ: Erlbaum.
- Berck, H. (1986). *Begriffe im Biologieunterricht. Versuch einer Entwirrung am Beispiel Sachkunde*, Köln: Aulis.
- Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model. Fundamental measurement in the human sciences*, Mahwah, NJ: Erlbaum.
- Borsboom, D., Mellenbergh, G.J. & van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, 110(2), 203–219.
- Bromme, R. & Bündler, W. (1994). Fachbegriffe und Arbeitskontext. Unterschiede in der Struktur chemischer Fachbegriffe bei verschiedenen Nutzergruppen. *Sprache und Kognition*, 13(4), 178–190.
- Bromme, R. & Rambow, R. (2001). Experten-Laien-Kommunikation als Gegenstand der Expertiseforschung. Für eine Erweiterung des psychologischen Bildes vom Experten. In Silbereisen, R.K. & Reitzle, M. (Hrsg.), *Psychologie 2000. Bericht über den 42. Kongress der Deutschen Gesellschaft für Psychologie in Jena 2000* (S. 541–550). Lengerich: Pabst Science Publishers.
- Buhlmann, R. & Fearn, A. (2000). *Handbuch des Fachsprachenunterrichts. Unter besonderer Berücksichtigung naturwissenschaftlich-technischer Fachsprachen*, Tübingen: Narr.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*, München: Pearson.
- Burmeister, C., Schmiemann, P. & Sandmann, A. (2010). *Learning with Biological Worked-out Examples*. Präsentation im Rahmen des Joint Meeting of Graduate Schools, Helsinki.
- Cassels, J.R.T. & Johnstone, A.H. (1985). *Words that matter in Science. A Report of a Research Exercise*, London: The Royal Society of Chemistry.
- Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychological Methods*, 3(3), 380–396.
- Embretson, S.E. & Reise, S. (2000). *Item Response Theory for Psychologists*, Mahwah, NJ: Erlbaum.
- Enrich, H. & Staeck, L. (Hrsg.) (1992). *Sprache und Verstehen im Biologieunterricht*, Alsbach/Bergstraße: Leuchtturm-Verlag.
- Fluck, H.-R. (1997). *Fachdeutsch in Naturwissenschaft und Technik. Einführung in die Fachsprachen und die Didaktik/Methodik des fachorientierten Fremdsprachenunterrichts (Deutsch als Fremdsprache)*, Heidelberg: Julius Groos.

- Göpferich, S. (1995). *Textsorten in Naturwissenschaften und Technik. Pragmatische Typologie – Kontrastierung – Translation*, Tübingen: Gunter Narr.
- Graf, D. (1989). *Begriffslernen im Biologieunterricht der Sekundarstufe I. Empirische Untersuchungen und Häufigkeitsanalysen*, Frankfurt am Main: Lang.
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In Beck, B. & Klieme, E. (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung*. DESI-Studie (S. 83–99). Weinheim: Beltz.
- Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In Hartig, J. & Klieme, E. (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (S. 17–36). Bonn, Berlin: Bundesministerium für Bildung und Forschung.
- Helmke, A. & Hosenfeld, I. (2004). Vergleichsarbeiten – Standards – Kompetenzstufen. Begriffliche Klärung und Perspektiven. In Wosnitza, M., Frey, A. & Jäger, R.S. (Hrsg.), *Lernprozess, Lernumgebung und Lerndiagnostik* (S. 56–75). Landau: Verlag Empirische Pädagogik.
- Kattmann, U. (1992). Von der Macht der Namen. Was mit biologischen Fachbegriffen gelernt wird. In Enrich, H. & Staeck, L. (Hrsg.), *Sprache und Verstehen im Biologieunterricht* (S. 91–118). Alsbach/Bergstraße: Leuchtturm-Verlag.
- Klein, W.P. (2003). Die Spannung zwischen Fach- und Gemeinsprache als Anlass für Sprachreflexion. *Deutschunterricht*, 56(2), 28–32.
- Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht. Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In Baumert, J., Bos, W. & Lehmann, R. (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe (S. 57–128). Opladen: Leske & Budrich.
- Klieme, E. & Steinert, B. (2004). Einführung der KMK-Bildungsstandards. *Der mathematische und naturwissenschaftliche Unterricht*, 57(3), 132–137.
- Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung. Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In Baumert, J., Bos, W. & Lehmann, R. (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit (S. 85–133). Opladen: Leske & Budrich.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*, Bonn: Bundesministerium für Bildung und Forschung.
- Klieme, E., Maag-Merki, K. & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In Hartig, J. & Klieme, E. (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (S. 5–15). Bonn, Berlin: Bundesministerium für Bildung und Forschung.
- Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, 52(6), 876–903.
- Knobloch, R., Sumfleth, E. & Walpuski, M. (2009). *Analysis of the quality of communication processes in small groups and the influence on the learning outcome*. Präsentation im Rahmen des Joint Meeting of Graduate Schools, Rovaniemi.
- Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) (2004). *Einheitliche Prüfungsanforderungen in der Abiturprüfung. Biologie*, München, Neuwied: Luchterhand.

- Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*, München, Neuwied: Luchterhand.
- Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*, München: Wolters Kluwer.
- Kramer, G. (2009). *Entwicklung und Überprüfung eines Strukturmodells der fachlichen Kommunikationskompetenz im Biologieunterricht*. Dissertation. Christian-Albrechts-Universität.
- Krischer, B. (2007). Hilfe - der Text ist zu schwer! Typische sprachliche Stolpersteine in Fachtexten – und wie man sie umgeht. *Lernchancen*, 10(59), 32–39.
- Kroß, A. & Lind, G. (2001). Einfluss des Vorwissens auf Intensität und Qualität des Selbsterklärens bei Lernen mit biologischen Beispielaufgaben. *Unterrichtswissenschaft*, 29(1), 5–24.
- Kulgemeyer, C. & Schecker, H. (2009). Kommunikationskompetenz in der Physik. Zur Entwicklung eines domänenspezifischen Kommunikationsbegriffs. *Zeitschrift für Didaktik der Naturwissenschaften*, 15, 131–153.
- Lewis, J. (2000). Genes, chromosomes, cell division and inheritance – do students see any relationship? *International Journal of Science Education*, 22(2), 177–195.
- Lienert, G.A. & Raatz, U. (1994). *Testaufbau und Testanalyse*, Weinheim: Beltz.
- Meloefski, R. (2007). Vom Alltagsbegriff zum Fachbegriff. *Der mathematische und naturwissenschaftliche Unterricht*, 60(4), 223–229.
- Moosbrugger, H. (2007). Item-Response-Theorie (IRT). In Moosbrugger, H. & Kelava, A. (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 215–259). Heidelberg: Springer.
- Moser, U., Ramseier, E., Keller, C. & Huber, M. (1997). *Schule auf dem Prüfstand. Eine Evaluation der Sekundarstufe I auf der Grundlage der „Third International Mathematics and Science Study“*, Chur: Rüegger.
- Neubrand, M. (2004). Der Prozess der Itementwicklung bei der nationalen Ergänzungsuntersuchung von PISA 2000. Vom theoretischen Rahmen zu den konkreten Aufgaben. In Neubrand, M. (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Nitz, S., Nerdel, C. & Prechtel, H. (2010). Language in science education and the influence of teachers' professional knowledge. In Tasar, M.F. & Cakmakci, G. (Hrsg.), *Contemporary science education research: pre-service and in-service teacher education* (S. 323–329). Ankara, Turkey: Pegem Akademi.
- Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest. Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, 30(1), S.120–135.
- Ramseier, E. (1997). *Naturwissenschaftliche Leistungen in der Schweiz. Vertiefende Analyse der nationalen Ergebnisse in TIMSS*, Bern: Amt für Bildungsforschung.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Nielsen und Lydiche.
- Roelcke, T. (2005). *Fachsprachen*, Berlin: Schmidt.
- Rohracher, H. (1988). *Einführung in die Psychologie*, München: Psychologie Verlags Union.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*, Bern: Hans Huber.
- Rudolph, G. (2007). „Diese Wörter kenne ich nicht.“ Fachbegriffe in Sachtexten erschließen. *Deutschunterricht*, 60(2), 20–25.
- Schaefer, G. (Hrsg.) (1980). *Kommunikative Grundlagen des naturwissenschaftlichen Unterrichts*, Weinheim: Beltz.
- Schecker, H. & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, 12, 45–66.
- Schmiemann, P. (2010). *Modellierung von Schülerkompetenzen im Bereich des biologischen Fachwissens*, Berlin: Logos.

- Shepardson, D.P. (1997). Of Butterflies and Beetles: First Graders' Ways of Seeing and Talking about Insect Life Cycles. *Journal of Research in Science Teaching*, 34(9), 873–889.
- Vollmer, G. (1980). *Sprache und Begriffsbildung im Chemieunterricht*, Frankfurt: Diesterweg.
- Walter, O. (2005). *Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten*, Lengerich: Pabst Science Publishers.
- Weinert, F.E. (2001). Concept of Competence. A Conceptual Clarification. In Rychen, D.S. & Salganik, L.H. (Hrsg.), *Defining and Selecting Key Competencies* (S. 45–65). Seattle, Toronto, Bern, Göttingen: Hogrefe.
- Weinert, F.E. (2002). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In Weinert, F.E. (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.
- Wellington, J.J. & Osborne, J. (2001). *Language and literacy in science education*, Buckingham: Open University Press.
- Wichter, S. (1994). *Experten- und Laienwortschätze. Umriss einer Lexikologie der Vertikalität*, Tübingen: Niemeyer.

KONTAKT

Philipp Schmiemann
Freie Universität Berlin
Fachbereich Biologie, Chemie, Pharmazie
Institut für Biologie
Didaktik der Biologie
Schwendenerstraße 1
14195 Berlin
philipp.schmiemann@fu-berlin.de

AUTORENINFORMATIONEN

Prof. Dr. Philipp Schmiemann
ist Juniorprofessor für Didaktik der Biologie an der Freien Universität Berlin. Seine Forschungsschwerpunkte sind Kompetenzmodellierung, Test- und Lernaufgaben im Fach Biologie.