Erschienen in: Meier, Monique; Wulff, Claudia; Zipprecht, Kathrin (Hrsg.), *Vielfältige Wege biologiedidaktischer Forschung. Vom Lernort Natur über Naturwissenschaftliche Erkenntnisgewinnung zur Lehrerprofessionalisierung; Festschrift für Prof. Dr. Jürgen Mayer*. Münster; New York: Waxmann, 2021, S. 65-75. DOI (Sammelwerk): 10.31244/9783830991182

Vanessa Fischer, Mariella Rothe[1], Elke Sumfleth,
Maik Walpuski & Nicole Wellnitz

Zur Konstruktion fächerübergreifend vergleichbarer Kompetenz-Testaufgaben

Wir haben Jürgen Mayer in den unterschiedlichsten Zusammenhängen als äußerst angenehmen Kooperationspartner kennengelernt, der immer ausgleichend und kompromisssuchend als kompetenter Moderator nach gemeinsamen Wegen gesucht hat. Wir entschuldigen uns schon im Voraus, dass er deswegen hier nicht als Ko-Autor auftreten kann. Sein Anteil – auch an diesem Beitrag – ist davon ganz unberührt. Wir danken Jürgen Mayer für jahrelange freundschaftliche Kooperation in verschiedenen Zusammenhängen.

1 Einleitung

In dem von der DFG geförderten Projekt IMBliCK (Einfluss von *I*nteresse und *M*otivation in den Fächern *Biologie* und *C*hemie auf Leistungsunterschiede in *K*ompetenztests) wurde der Einfluss affektiver Faktoren, wie Interessantheit der Aufgaben und motivationale Anregung, auf die Bearbeitung von Leistungstestaufgaben in den Fächern Biologie und Chemie untersucht. Dazu war es notwendig, vergleichbare Testinstrumente zu entwickeln, die zum einen die Kompetenzen von Schülerinnen und Schülern in beiden Fächern messbar machen und zum anderen vergleichbare Anforderungen aufweisen, um den Einfluss der affektiven Faktoren systematisch untersuchen zu können. Die besondere Herausforderung lag darin, Kompetenztestaufgaben zu entwickeln, die hinsichtlich ihrer Schwierigkeit über die Fächer Biologie und Chemie vergleichbar sind. Ziel war es daher, zunächst konkrete Kriterien zu entwickeln, die bei der Erstellung der Testaufgaben beachtet werden sollten. Bereits bekannt war, dass der Kontext einen bedeutsamen Einfluss auf die Interessantheit von Aufgaben hat. Zudem zeigte sich bisher, dass die Komplexität einer Aufgabe und der zu erbringende kognitive Prozess einen Einfluss auf die Aufgabenschwierigkeit haben. Unter Berücksichtigung dieser drei Parameter wurden in diesem Projekt kontextorientierte Kompetenztestaufgaben mit vergleichbaren Merkmalsausprägungen entwickelt. Im Fokus dieses Beitrags steht die Vergleichbarkeit der entwickelten Testaufgaben.

1 geb. Mariella Roesler

2 Ablaufplan zur Testkonstruktion

Terzer, Hartig und Upmeier zu Belzen (2013, S. 53) schlagen auf Basis der Literaturlage einen siebenschrittigen Ablaufplan für die Konstruktion von Kompetenztests vor: „(1) Formulierung der theoretischen Fundierung, (2) Testkonzeption, (3) Systematisierung der Itemkonstruktion, (4) Entwicklung einer Konstruktionsanleitung, (5) Itementwicklung, (6) Itemerprobung und -selektion und (7) Festlegung des Erhebungsdesigns". Diese Schritte sollten nacheinander durchlaufen werden, wobei einzelne Schritte ggf. auch wiederholt werden können.

Basis für die Konstruktion eines Tests ist die Definition des Untersuchungsgegenstands (Wilson, 2005), also die *Formulierung der theoretischen Fundierung*. So sind bei Kompetenzmessungen z. B. Ausführungen zum verwendeten Kompetenzbegriff unumgänglich. Hinzu kommen ggf. weitere zu berücksichtigende Aspekte, die entweder mit demselben Testinstrument oder mit weiteren Instrumenten erhoben werden. Die spezifische Konkretisierung der Ausführungen führt zu definierten Operationalisierungen bzw. Kompetenzmodellen als Grundlagen für die Entwicklung der Testitems. Der Schritt der *Testkonzeption* basiert auf der Festlegung des Untersuchungsziels (Jonkisz, Moosbrugger & Brandt, 2012; Neuhaus & Braun, 2007), z. B. der Überprüfung eines Kompetenzmodells und der Auswahl der Messmethode(n) unter Berücksichtigung institutioneller und anderer Rahmenbedingungen. Im Rahmen der *Systematisierung der Itemkonstruktion* wird in einem ersten Schritt mit Hilfe eines Manuals zur Itementwicklung sichergestellt, dass die Operationalisierung der Kompetenzen oder Kompetenzausprägungen eindeutig und nachvollziehbar ist (Köller, 2008). Dies kann auf Basis von Indikatoren geschehen, die über ein Expertenrating abgesichert werden können (Hartig & Jude, 2007; Rost, 2004). Diese Entscheidungen werden in *Konstruktionsanleitungen* festgehalten, damit Aufgaben z. B. auch von verschiedenen Personen entwickelt werden können (Hartig & Jude, 2007; Wilson, 2005). In diesen Anleitungen werden auch die Struktur der Aufgaben sowie allgemeine Angaben zum Aufgabenstamm, zur Aufgabenstellung und zu den Antwortmöglichkeiten festgehalten. Zum einen muss die formale Gestaltung der Items, wie das Antwortformat (Jonkisz et al., 2012), die Textlänge und die Art der Abbildungen (Jonkisz et al., 2012; Rost, 2004) definiert werden. Zum anderen spielen inhaltliche Merkmale wie der angestrebte (Bybee, 2002; Walpuski & Ropohl, 2014) in Abhängigkeit von der Zielgruppe und mit Blick auf die notwendige Wissensbasis (Prenzel, Häußler, Rost & Senkbeil, 2002) bzw. curriculare Passung und den inhaltlichen Geltungsbereich eine wichtige Rolle. Dabei ist auch die Auswahl der Kontexte (Neuhaus & Braun, 2007; Hammann, 2006) zu be-

rücksichtigen. Basierend auf der Konstruktionsanleitung erfolgt die *Itementwicklung*. Idealerweise werden auf Basis curricularer Anforderungen und bekannter Schülervorstellungen bzw. Verständnisprobleme zunächst offene Items entwickelt und diese dann anhand der Schülerantworten in Multiple-Choice-Aufgaben überführt (Wilson, 2005), um plausible Distraktoren zu erhalten. Dabei sind ähnliche Formulierungen aller Antwortalternativen essentiell, um Anhaltspunkte für die richtige Antwortalternative zu vermeiden. Unterschiedliche Testhefte sind leichter zusammenzustellen, wenn logische Abhängigkeiten zwischen Items vermieden werden (Haladyna, 1999). Außerdem müssen die Aufgabentexte sprachlich einfach gestaltet sein (Jonkisz et al., 2012; Lienert & Raatz, 1998), um den Einfluss der Sprachkompetenz auf das Testergebnis gering zu halten. Die Anwendung der Methode des lauten Denkens oder der Kommentierung der Aufgaben bei der Bearbeitung durch die Zielgruppe kann zusätzlich zur Validitätsprüfung herangezogen werden. Für die Pilotierung der Items sollten grundsätzlich mehr Items konstruiert werden als später benötigt werden, um unbefriedigende Items streichen zu können. Neben der Testart ist eine ausreichende Itemzahl zur Abdeckung der interessierenden Aspekte wichtig, damit eine Schätzung der statistischen Parameter mit hinreichender Reliabilität möglich ist.

3 Instrumentenentwicklung

Für das im Projekt IMBliCK herangezogene Kompetenzkonstrukt bilden die Fächer Biologie und Chemie die inhaltliche Domäne. Die in den Bildungsstandards für diese beiden Fächer festgelegten Kompetenzbereiche (Fachwissen, Erkenntnisgewinnung, Kommunikation und Bewertung) und die dort formulierten Kompetenzen orientieren sich an der kognitiven Facette der Weinert'schen Kompetenzdefinition (Weinert, 2001). In Anbetracht der Fragestellung des Projekts IMBliCK zum Einfluss affektiver Faktoren auf die Bearbeitung von Aufgaben in den Fächern Biologie und Chemie wurden zwei Kompetenzbereiche ausgewählt, die sich in ihrer Interessantheit möglichst stark voneinander unterscheiden. Nach Holstermann und Bögeholz (2007) sind dies Aufgaben in den Kompetenzbereichen Bewertung und Fachwissen, wobei jene zur Bewertung ein höheres Interesse erzeugen als solche zum Fachwissen. Eine tiefergehende Betrachtung der beiden Kompetenzbereiche zeigt eine Vielzahl an Gemeinsamkeiten zwischen den Bezugsdisziplinen (Tab. 1), sodass eine vergleichbare Konstruktion von Testaufgaben möglich ist.

Tabelle 1: Kompetenzbereiche der Fächer Biologie (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005a, S. 7) und Chemie (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005b, S. 7)

Fächer	**Kompetenzbereiche**	
	Fachwissen	Bewertung
Biologie	Lebewesen, biologische Phänomene, Begriffe, Prinzipien, Fakten kennen und den Basiskonzepten zuordnen	Biologische Sachverhalte in verschiedenen Kontexten erkennen und bewerten
Chemie	Chemische Phänomene, Begriffe, Gesetzmäßigkeiten kennen und Basiskonzepten zuordnen	Chemische Sachverhalte in verschiedenen Kontexten erkennen und bewerten

Der Kompetenzbereich Fachwissen umfasst das Kennen von Phänomenen und Begriffen sowie Prinzipien bzw. Gesetzmäßigkeiten aus den entsprechenden Bezugsdisziplinen. Kauertz, Fischer, Mayer, Sumfleth und Walpuski (2010) zeigen, dass Theorien aus den entsprechenden Bezugsdisziplinen zur Lösung von naturwissenschaftlichen Problemen eingesetzt werden und zudem relevante Prinzipien, Modelle und Konzepte dazu angewendet werden. Der Kompetenzbereich Bewertung beinhaltet das Einbinden von fachlichen Informationen in einen Entscheidungsprozess sowie die Reflexion solcher Entscheidungsprozesse (Hostenbach et al., 2011; Kauertz et al., 2010).

Die zu Beginn des Projekts erstellte Aufgabenkonstruktionsanleitung ermöglichte eine systematische Konstruktion von Items in beiden Fächern zu beiden Kompetenzbereichen. Zur Operationalisierung der zu überprüfenden Kompetenzen wurde ein bestehendes Kompetenzmodell verwendet (ESNaS-Modell: Walpuski et al., 2010), um hinsichtlich ihrer erwarteten Schwierigkeit vergleichbare Items für die Fächer zu entwickeln. Dazu wurden zunächst Aufgabenmerkmale definiert, die die Schwierigkeit beeinflussen. Bisherige Befunde zeigen, dass die Komplexität einer Aufgabe und der zu erbringende kognitive Prozess, das Aufgabenformat (Walpuski & Ropohl, 2014) sowie ggf. vorhandene Abbildungen (Hartmann, 2013) einen Einfluss auf die Aufgabenschwierigkeit haben. Diese Merkmale einer Aufgabe wurden in dem Projekt über beide Fächer und beide Kompetenzbereiche konstant gehalten. Zudem wurden vier Kontexte (Gesundheit, Umwelt, Technik, Natürliche Ressourcen) gewählt, um eine möglichst hohe Varianz in der Interessantheit und motivationalen Anregung in den Aufgaben zu erhalten. Dies bedeutet, dass die Aufgaben in beiden Fächern zu beiden Kompetenzbereichen parallel entwickelt wurden, um so möglichst vergleichbare Testinstrumente zu erhalten. So wurden zu jedem Kontext 32 Items pro Fach entwickelt, die sich zu gleichen Teilen auf beide Kompetenzbereiche verteilen. Insgesamt wurden demnach

256 Items konstruiert. Die Items wurden alle für eine mittlere Komplexitätsstufe (ein Zusammenhang) erstellt. Um die Varianz in der Aufgabenschwierigkeit zu erhöhen, wurden die kognitiven Prozesse in den Items variiert (selegieren, organisieren, integrieren). Da die Durchführung der Testungen in zwei benachbarten Bundesländern (Hessen, NRW) stattfand, konnten die länderspezifischen Kernlehrpläne bzw. Kerncurricula nicht als Grundlage für die inhaltliche Ausgestaltung der Items dienen. Es wurden daher die Bildungsstandards hinzugezogen, die länderübergreifend gültig sind. Da die Messung der Kompetenzen durch einen Paper-Pencil-Test erfolgte, wurden die Items im Multiple-Choice Single-Select und im offenen Format zu gleichen Anteilen für beide Fächer und Kompetenzbereiche entwickelt. Somit sollte durch das Antwortformat zusätzliche Varianz in den Aufgabenschwierigkeiten erzeugt werden. Abbildungen wurden nur eingesetzt, wenn sie funktional und von daher bedeutsam für die Beantwortung der Fragestellung waren. Die Qualität der Items wurde hinsichtlich ihrer Einstufung in das ESNaS-Modell und ihrer fachlichen Richtigkeit durch mehrere Korrekturschleifen innerhalb des Projekts und in Kooperation mit weiteren Fachdidaktikern sichergestellt.

Die entwickelten Items zu beiden Fächern und Kompetenzbereichen wurden auf Grundlage der Kontextzugehörigkeit zu Aufgaben geclustert (insgesamt 8 Aufgaben pro Fach und Kompetenzbereich). Jede Aufgabe besteht aus vier Items, die einem Fach, einem Kompetenzbereich und einem Kontext zugeordnet werden können. Nach der Bearbeitung der Aufgaben machten die Schülerinnen und Schüler Angaben zur Interessantheit (Haugwitz, 2009) und zur motivationalen Anregung (Boekaerts, 2002; Sundre, 2007) der Aufgaben, sodass die Einflüsse dieser affektiven Faktoren auf die Aufgabenschwierigkeit analysiert werden konnten.

4 Untersuchungsfragen und Hypothesen

Ziel der Arbeitsschritte im Projekt IMBliCK war es, Kompetenztestaufgaben in den Fächern Biologie und Chemie zu entwickeln, die hinsichtlich ihrer erwarteten Aufgabenschwierigkeiten vergleichbar sind, um so systematisch den Einfluss affektiver Faktoren bei der Bearbeitung von Kompetenztestaufgaben zu untersuchen. Dabei war es wichtig, Items zu entwickeln, die hinsichtlich spezifischer Aufgabenmerkmale vergleichbar sind, jedoch die Besonderheiten der entsprechenden Fachdisziplin nicht außer Acht lassen. So ist es in dem Projekt IMBliCK eine besondere Herausforderung gewesen, die spezifischen Merkmale der einzelnen Fachdisziplinen zu konkretisieren sowie Aufgabenmerkmale für die Konstruktion zu definieren, die eine vergleichbare Konstruktion er-

möglichten. Zur Überprüfung der Testqualität war folgende Forschungsfrage handlungsleitend:

> *Lassen sich die neu entwickelten Testinstrumente hinsichtlich der Fachzugehörigkeit und des Kompetenzbereichs empirisch trennen?*

Erwartbar waren vier verschiedene Modelle, die durch die folgenden Hypothesen beschrieben werden:

1. Eindimensionales Modell: Die Items zu den Kompetenzbereichen Fachwissen und Bewertung für die Fächer Biologie und Chemie sind einer gemeinsamen Skala zuzuordnen. Aufgrund gemeinsamer theoretischer Annahmen können die Items nicht als trennbare Skalen abgebildet werden.
2. Zweidimensionales Modell I: Die Items zu den Fächern Biologie und Chemie bilden zwei trennbare Skalen. Trotz gemeinsamer theoretischer Annahmen können diese beiden Konstrukte empirisch voneinander getrennt erfasst werden.
3. Zweidimensionales Modell II: Die Items zu den Kompetenzbereichen Fachwissen und Bewertung bilden zwei trennbare Skalen. Trotz gemeinsamer theoretischer Annahmen können diese beiden Skalen empirisch voneinander getrennt erfasst werden.
4. Vierdimensionales Modell: Die Items zu den Fächern Biologie und Chemie und den jeweiligen Kompetenzbereichen Fachwissen und Bewertung bilden vier trennbare Skalen, die trotz gemeinsamer theoretischer Annahmen empirisch voneinander getrennt erfasst werden können.

Um Unterschiede hinsichtlich des Fachs und des Kompetenzbereichs sinnvoll untersuchen zu können, sollte der Test nach Möglichkeit alle vier Dimensionen getrennt abbilden können (Hypothese 4).

5 Qualität der Testinstrumente

Für die Analyse der Qualität des entwickelten Testinstruments wurden Rasch-Analysen mit ConQuest® (Wu, Adams, Wilson & Haldane, 2007) durchgeführt. Dazu wurden zunächst die Items getrennt nach Fächern und Kompetenzbereichen hinsichtlich ihrer Qualität überprüft. Im Anschluss wurden alle Items mit guter Modellpassung ($0.8 < MNSQ < 1.25$) in die weiteren Analysen mit einbezogen.

Im nächsten Schritt wurden ein eindimensionales Modell, zwei zweidimensionale Modelle und ein vierdimensionales Modell berechnet, um die Hypothesen zu überprüfen. Dabei wurden die Werte der berechneten Modell-

parameter (Deviance) miteinander verglichen und hinsichtlich statistischer Signifikanz geprüft. Dabei zeigte sich, dass das vierdimensionale Modell am besten zu den Daten passt (ΔDeviance $p < .001$), sodass davon auszugehen ist, dass es sich bei den Skalen Fachwissen Biologie, Bewertungskompetenz Biologie, Fachwissen Chemie und Bewertungskompetenz Chemie um vier empirisch trennbare Konstrukte handelt, die getrennt voneinander erfasst werden können. So zeigte sich, dass es trotz der vergleichbaren Konstruktion der Items zwischen den Fächern und Kompetenzbereichen neben den kognitiven Merkmalen, die die Schwierigkeit der Items beeinflussen, weitere Merkmale zu geben scheint, die dafür sorgen, dass es messbare Unterschiede in den Aufgaben zwischen den Fächern und Kompetenzbereichen gibt. Somit konnten im Anschluss Analysen für die einzelnen Skalen getrennt voneinander vorgenommen werden. In einem nächsten Schritt wurden die latenten Korrelationen untersucht, um die Stärke des Zusammenhangs zwischen den einzelnen Skalen zu prüfen (Tab. 2).

Tabelle 2: Latente Korrelationen der Itemschwierigkeiten zwischen den Fächern Biologie und Chemie und den Kompetenzbereichen Fachwissen und Bewertung

	Biologie Bewertung	**Chemie Fachwissen**	**Chemie Bewertung**
Biologie Fachwissen	.79	.85	.72
Biologie Bewertung		.72	.84
Chemie Fachwissen			.64

Die latente Korrelation zwischen den Fächern, unabhängig vom Kompetenzbereich, liegt in einem hohen Bereich ($r = .91$), ähnlich wie bei Rost, Walter, Carstensen, Senkbeil und Prenzel (2004) berichtet wird. Bei genauerer Betrachtung der latenten Korrelationen zwischen den Fächern, abhängig vom Kompetenzbereich, zeigt sich, dass die Korrelationen insgesamt in einem mittleren bis hohen Bereich liegen. Dies ist ein Hinweis darauf, dass es sich bei den Skalen um verwandte Konstrukte handelt. Die höchsten Korrelationen bestehen innerhalb desselben Kompetenzbereichs zwischen den Aufgaben aus beiden Fächern. Dieses Ergebnis weist daraufhin, dass die gewählten Aufgaben bereichsspezifische Kompetenzen messen. Auch kann man feststellen, dass die latenten Korrelationen innerhalb der Fächer bedeutsam sind. Während im Fach Biologie die Korrelationen mit .79 in einem hohen Bereich liegen, bewegen sie sich im Fach Chemie mit .72 in einem mittleren Bereich. Dies zeigt, dass bei der Konstruktion der Testinstrumente vergleichbare Aufgaben zwischen den Fächern entwickelt wurden und ist ein Beleg für die konvergente Validität der entwickelten Testinstrumente.

In einem nächsten Schritt wurden die Aufgaben hinsichtlich ihrer Schwierigkeiten zwischen den Fächern und Kompetenzbereichen verglichen. Dabei zeigte sich, dass sich die Aufgabenschwierigkeiten zwischen den Fächern, unabhängig vom Kompetenzbereich, nicht signifikant voneinander unterscheiden ($t(51.648) = -1.84$, $p = .072$, $d = 0.48$). Dagegen unterscheiden sich die Aufgabenschwierigkeiten zwischen den Kompetenzbereichen, unabhängig vom Fach, signifikant mit einem großen Effekt ($t(62) = -5,.46$, $p < .001$, $d = 1.36$). Aufgaben aus dem Kompetenzbereich Fachwissen sind signifikant schwerer als Aufgaben aus dem Kompetenzbereich Bewertung. Dieser Unterschied ist auf die Aufgaben im Fach Chemie zurückzuführen, wohingegen sich die Aufgaben im Fach Biologie nicht signifikant zwischen den Kompetenzbereichen unterscheiden ($t(30) = 1.73$, $p = .095$, $d = 0.61$). Trotz der Kontrolle bekannter schwierigkeitsbestimmender Aufgabenmerkmale sind Aufgaben zum Kompetenzbereich Bewertung im Fach Chemie im Vergleich signifikant leichter ($t(30) = -6.97$, $p < .001$, $d = 2.47$). Dies deutet darauf hin, dass es neben den bereits bekannten schwierigkeitserzeugenden Merkmalen weitere Merkmale gibt, die für die Schwierigkeit einer Aufgabe von Bedeutung sind. Zudem ist die Dimensionsanalyse ein weiteres Indiz dafür, dass es Faktoren gibt, die die systematisch parallel entwickelten Aufgaben unterscheidbar zu machen scheinen.

6 Zusammenfassung

Die Analyse der neu entwickelten Testinstrumente zeigt, dass es sich um vier empirisch trennbare Skalen handelt, die getrennt voneinander betrachtet werden können. Trotzdem zeigen sich zwischen den Skalen hohe Korrelationen, die unter anderem auf die systematische Konstruktion schwierigkeitsbestimmender Aufgabenmerkmale zurückzuführen sind. Die Vergleiche der Aufgabenschwierigkeiten zeigen zudem, dass es im Fach Chemie Unterschiede in den Aufgabenschwierigkeiten zwischen den Kompetenzbereichen gibt, trotz der Kontrolle schwierigkeitsbestimmender Merkmale.

Die Ergebnisse verdeutlichen, dass neben den bekannten schwierigkeitsbestimmenden kognitiven Aufgabenmerkmalen andere Faktoren einen Einfluss auf die Aufgabenschwierigkeit haben. Dies können inhaltliche Unterschiede oder affektive Faktoren sein, die für die unterschiedlichen Skalen verantwortlich sind. Diese Unterschiede in den Skalen werden zudem deutlich, wenn man die Aufgabenschwierigkeiten betrachtet. Dabei fällt auf, dass Aufgaben zum Kompetenzbereich Bewertung im Fach Chemie leichter zu lösen sind als Aufgaben zum gleichen Kompetenzbereich im Fach Biologie und auch im Vergleich zu Aufgaben zum Kompetenzbereich Fachwissen in beiden Fächern.

Dies lässt sich nicht auf Aufgabenmerkmale zurückführen, die in bisherigen Studien als schwierigkeitsbestimmend erkannt wurden, sondern vermutlich auf motivationale Faktoren.

Anmerkung

Das diesem Artikel zugrundeliegende Vorhaben wurde von der DFG im Rahmen einer Sachbeihilfe gefördert (MA 1792/6–1, SU 187/12–1 WA 2829/5–1).

Literatur

Boekaerts, M. (2002). The On-Line Motivation Questionnaire: A self-report instrument to assess students' context sensitivity. *New Directions in Measures and Methods, 12*, 77–120.

Bybee, R.W. (2002). Scientific Literacy – Mythos oder Realität? In W. Gräber, P. Nentwig, T. Koballa & R. Evans (Hrsg.), *Scientific Literacy. Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung* (S. 21–43). Opladen: Leske & Budrich.

Haladyna, T.M. (1999). *Developing and Validating Multiple-Choice Test Items* (2. Auflage). Mahwah, NJ: Lawrence Erlbaum Associates.

Hammann, M. (2006). Kompetenzförderung und Aufgabenentwicklung. *Der mathematisch-naturwissenschaftliche Unterricht*, *59*(2), 85–95.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 17–36). Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Hartmann, S. (2013). *Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz.* Dissertation: Universität Duisburg-Essen.

Haugwitz, M. (2009). *Kontextorientiertes Lernen und Concept Mapping im Fach Biologie: Eine experimentelle Untersuchung zum Einfluss auf Interesse und Leistung unter Berücksichtigung von Moderationseffekten individueller Voraussetzungen beim kooperativen Lernen.* Dissertation: Universität Duisburg-Essen.

Holstermann, N. & Bögeholz, S. (2007). Interesse von Jungen und Mädchen an naturwissenschaftlichen Themen am Ende der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, 13, 71–86.

Hostenbach, J., Fischer, H.E., Kauertz, A., Mayer, J., Sumfleth, E. & Walpuski, M. (2011). Modellierung der Bewertungskompetenz in den Naturwissenschaften zur Evaluation der Nationalen Bildungsstandards. *Zeitschrift für Didaktik der Naturwissenschaften, 17,* 261–288.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27–74). Berlin: Springer.

Kauertz, A., Fischer, H.E., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 135–153.

Köller, O. (2008). Bildungsstandards – Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. *Zeitschrift für Pädagogik, 54*(2), 163–173.

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Neuhaus, B. & Braun, E. (2007). Testkonstruktion und Testanalyse – praktische Tipps für empirisch arbeitende Didaktiker und Schulpraktiker. In H. Bayrhuber, D. Elster, D. Krüger & H.J. Vollmer (Hrsg.), *Kompetenzentwicklung und Assessment* (S. 135–164). Innsbruck: StudienVerlag.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30*(2), 120–135.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.

Rost, J., Walter, O., Carstensen, C.H., Senkbeil, M. & Prenzel, M. (2004). Naturwissenschaftliche Kompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 111–146). Münster: Waxmann.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005a). *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München, Neuwied: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005b). *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München, Neuwied: Luchterhand.

Sundre, D.L. (2007). *The Student Opinion Scale (SOS): A measure of examinee motivation. Test Manual.* Harrisonbourg, VA: The Center for Assessment & Research Studies.

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.

Walpuski, M., Kauertz, A., Kampa, N., Fischer, H.E., Mayer, J., Sumfleth, E. & Wellnitz, N. (2010). ESNaS – Evaluation der Standards für die Naturwissenschaften in der Sekundarstufe I. In A. Gehrmann, U. Hericks & M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle – Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 171–184). Bad Heilbrunn: Klinkhardt.

Walpuski, M. & Ropohl, M. (2014). Statistische Verfahren für die Analyse des Einflusses von Aufgabenmerkmalen auf die Schwierigkeit. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 385–398). Berlin: Springer.

Weinert, F.E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F.E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M.L., Adams, R.J., Wilson, M.R. & Haldane, S.A. (2007). *Acer ConQuest. Version 2.0. Generalised Item Response Modelling Software.* Camberwell: Australian Council for Educational Research.

DOI: 10.17185/duepublico/77026

URN: urn:nbn:de:hbz:465-20221013-095452-5

Fischer, V.; Rothe, M.; Sumfleth, E.; Walpuski, M.; Wellnitz, N.: Zur Konstruktion fächerübergreifend vergleichbarer Kompetenz-Testaufgaben. In: Meier, M.; Wulff, C.; Zipprecht, K. (Hrsg.), *Vielfältige Wege biologiedidaktischer Forschung. Vom Lernort Natur über Naturwissenschaftliche Erkenntnisgewinnung zur Lehrerprofessionalisierung; Festschrift für Prof.Dr. Jürgen Mayer*. Münster; New York: Waxmann, 2021, S. 65-75.