Z Erziehungswiss (2015) 18:763–779
DOI 10.1007/s11618-015-0646-2

ALLGEMEINER TEIL

ZfE

Analyse der Aufgaben zur Evaluation der Bildungsstandards in Physik - Differenzierung von schriftsprachlichen Fähigkeiten und Fachlichkeit

Hendrik Härtig · Patricia Heitmann · Jan Retelsdorf

Online publiziert: 9. Juli 2015

Zusammenfassung Bei der Evaluation der Bildungsstandards in den naturwissenschaftlichen Fächern sieht das Itemdesign vor, benötigtes Vorwissen innerhalb der Testung vorzugeben und damit auch im Sinne einer Kompetenzorientierung *Umgang mit Fachwissen* zu evaluieren. Dies setzt voraus, dass die Schülerinnen und Schüler die Informationen angemessen rezipieren. Es wird daher angenommen, dass schriftsprachliche Fähigkeiten eine zentrale Rolle für die Erfassung physikbezogener Kompetenzen in Leistungstests spielen, zumal einige Items schriftliche Antworten verlangen. Der vorliegende Beitrag geht daher der Frage nach, in welchem Umfang schriftsprachliche Fähigkeiten im Zusammenhang mit den Ergebnissen der Evaluation der Bildungsstandards im Kompetenzbereich *Umgang mit Fachwissen* des Fachs Physik steht. Die Studie zeigt an einer Stichprobe von $N=1961$ Auszubildenden mittels Vergleich logistischer Modelle auf, dass die Items der Bildungsstandards von schriftsprachlichen Fähigkeiten beeinflusst werden. Ferner wiesen Items mit offenem Format im Mittel eine höhere Itemschwierigkeit auf im Vergleich zu Items mit geschlossenem Format, wobei sich eine Interaktion zwischen Antwortformat und schriftsprachlichen Fähigkeiten andeutet.

Die Daten für die hier präsentierten Analysen stammen aus dem Projekt ManKobE („Mathematisch-naturwissenschaftliche Kompetenzen in der beruflichen Erstausbildung“), das aus Mitteln der Leibniz-Gemeinschaft gefördert wird (Förderkennzeichen: SAW−2012-IPN−2).

Prof. Dr. H. Härtig (✉) · Prof. Dr. J. Retelsdorf
Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik Kiel (IPN),
Olshausenstraße 62,
24118 Kiel, Deutschland
E-Mail: haertig@ipn.uni-kiel.de

Prof. Dr. J. Retelsdorf
E-Mail: jretelsdorf@ipn.uni-kiel.de

Dr. P. Heitmann
Institut zur Qualitätsentwicklung im Bildungswesen (IQB),
Luisenstr. 56,
10117 Berlin, Deutschland
E-Mail: patricia.heitmann@iqb.hu-berlin.de

Schlüsselwörter Evaluation der Bildungsstandards · Kompetenztests · Schriftsprachliche Fähigkeiten · Antwortformate

Analyses of the tasks for evaluating the educational standards in physics – Differentiation between written language proficiency and content knowledge

Abstract Within the evaluation of National Educational Standards in science education all items provide information required for the answer. The items thereby are also competency-based. We assume that general language proficiency plays a central role in order to assess achievement in physics since some answers require an open response. Thus, this paper addresses the question to what extent general language proficiency can affect the results of the evaluation of the educational standards in physics education for the competence *content knowledge*. Drawing on a sample of $N = 1961$ vocational trainees first, it is shown that the National Educational Standards items are affected by general language proficiency to a relevant degree. Furthermore, items in open response format are more difficult than items in a closed response format, with a slight interaction between response format and general language proficiency.

Keywords Answer format · Competence tests · Evaluation of educational standards · Written language proficiency

1 Einleitung

Die Bildungsstandards für den mittleren Schulabschluss in Physik beschreiben seit nunmehr einer Dekade das intendierte Ziel des Physikunterrichts in der Sekundarstufe I (KMK 2003, 2005). Im Jahr 2012 wurde vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) ein repräsentativer Ländervergleich durchgeführt, der sowohl der Erfassung des Leistungsstandes der Lernenden als auch der empirischen Validierung der formulierten Standards diente (Pant et al. 2013). Die eingesetzten Aufgaben wurden basierend auf einem Kompetenzstrukturmodell von Lehrkräften sowie Vertreterinnen und Vertretern aus Ministerien beziehungsweise Landesinstituten aus den Bundesländern entwickelt. Fokus der Aufgaben sind die fachbezogenen Kompetenzen in Physik. Der Kompetenzbereich Fachwissen umfasst der Intention der Bildungsstandards folgend den aktiven Umgang mit dem Fachwissen zum Lösen fachlicher Probleme. Lernende sollen mit Hilfe einer strukturierten Wissensbasis Problemstellungen zu physikalischen Phänomenen, Begriffen oder Gesetzmäßigkeiten bewältigen können (KMK 2005). Allerdings lassen sich andere kognitive, motivationale und volitionale Personenmerkmale als Varianzanteile der Performanz im Rahmen der Testung nicht ausschließen.

Im Rahmen der Kompetenzdefinition von Weinert (2001) und der davon ausgehenden Operationalisierung innerhalb der Evaluation der Bildungsstandards wird dies mitunter bewusst berücksichtigt. Insbesondere beim Kompetenzbereich Fach-

wissen zeigt sich, dass über alle Schulformen und Bundesländer hinweg die Lehrpläne kaum gemeinsame Inhalte beschreiben. Daher wurde entschieden, benötigtes Vorwissen innerhalb der Testung jeweils im Stamm eines Items vorzugeben und damit auch im Sinne einer Kompetenzorientierung *Umgang mit Fachwissen* zu evaluieren (Pant et al. 2013). Dies impliziert allerdings, dass die Schülerinnen und Schüler in der Lage sind, die Informationen angemessen zu rezipieren. Damit ergibt sich vermutlich zumindest eine Überlagerung der physikbezogenen Kompetenzen und (schrift)sprachlich[1] rezeptiven Fähigkeiten für alle Aufgaben. Zusätzlich erfordern manche Aufgaben schriftliche Antworten, was unter Umständen zusätzlich sprachlich produktive Fähigkeiten erfordert. Während andere Personenmerkmale über die Items hinweg durch Variation der Itemcharakteristika in Teilen kontrolliert werden können (z. B. mathematische Kompetenzen oder konvergentes Denken), ist dies bei den sprachlichen Fähigkeiten nicht der Fall (vgl. Kauertz et al. 2010; Ropohl 2010).

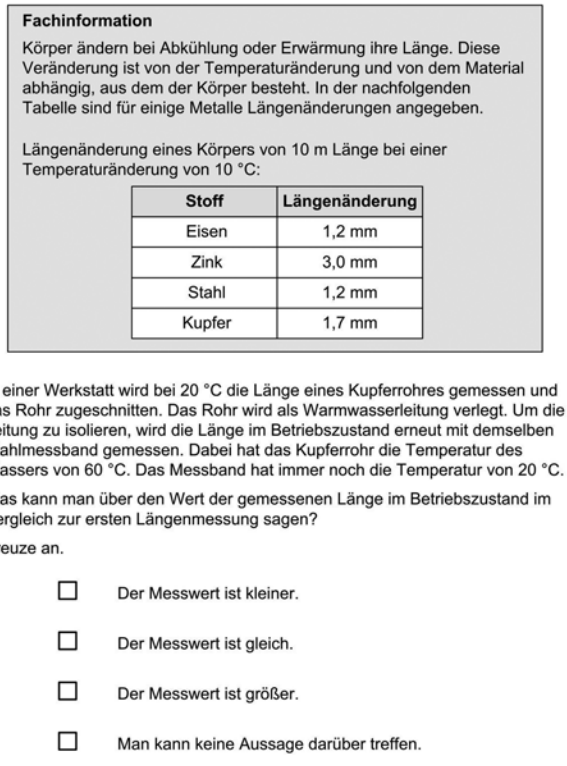
Zusammengefasst wird vermutet, dass die schriftliche Modalität sprachlicher Fähigkeiten (vgl. Jude et al. 2008, S. 192) Einfluss auf die Performanz im Bereich *Umgang mit Fachwissen* in der Evaluation der Bildungsstandards hat. In dem vorliegenden Beitrag werden in einer Untersuchung daher die drei Bereiche *Umgang mit Fachwissen Physik, sprachliche Fähigkeiten* und *Physik Fachwissen* aufeinander bezogen.

2 Theoretischer Hintergrund

Es wird davon ausgegangen, dass *sprachliche Fähigkeiten* eine wesentliche Variable für akademischen Erfolg im Allgemeinen, aber auch speziell für das Lernen in den Naturwissenschaften sind (vgl. Härtig et al. 2015). Letzteres spiegelt sich unter anderem in hohen korrelativen Zusammenhängen zwischen sprachlichen Fähigkeiten und Leistungen in naturwissenschaftlichen Tests wider (O'Reilly und McNamara 2007; Cromley 2009). In der Forschung wurde dabei in der Vergangenheit vor allem das Lesen naturwissenschaftsbezogener Texte als Teil des Lernens empirisch untersucht, wobei schriftsprachlich rezeptive Fähigkeiten im Fokus stehen, zum Beispiel im Rahmen von Untersuchungen zum Textverständnis. Das Textverständnis ist ein komplexer Prozess, der im Sinne von van Dijk und Kintsch (1983) als Bildung mentaler Repräsentationen interpretiert werden kann. Die Lesenden nehmen den Inhalt auf der Basis ihres Vorwissens wahr, interpretieren ihn und bilden daraus neue Wissenselemente (Norris und Phillips 2003). Dies scheint zu Interaktionen zwischen Personen- und Textmerkmalen zu führen, wobei sich zum Beispiel Veränderungen bei der Text-Bild Interaktion sowohl lernförderlich als auch lernhinderlich auswirken können (Sumfleth und Tiemann 2000; Schmeck 2011). Im Fokus dieses Artikels steht die Sichtweise der psychologisch orientierten Textverständnisforschung und der Nomenklatur von Jude et al. (2008): Als Fähigkeiten der Schülerinnen und Schüler lassen sich unter anderem sprachliche und fachliche Fähigkeiten beschreiben. Die fachlichen Fähigkeiten werden in diesem Artikel als *Fachwissen Physik* bezeichnet. Die sprachlichen Fähigkeiten unterteilen sich in die auditive und die schriftliche Modalität, sowie die Prozesse Produktion und Rezeption, wobei im Folgenden nur die schriftliche Modalität als *sprachliche Fähigkeiten* berücksichtigt wird.

Die benötigten sprachlichen Fähigkeiten bei der Bearbeitung von Testitems reichen von der Identifikation von Buchstaben und Entschlüsselung einzelner Wortbedeutungen bis zur Herstellung von Beziehungen zwischen mehreren Sätzen (Francis et al. 2006; Perfetti 2007; Hall et al. 2015). Für den Verstehensprozess als Interaktion spielt neben dem Wissen über Wortbedeutungen, und dem Vermögen Schlussfolgerungen aus zu verknüpfenden Textteilen zu ziehen, auch das Vorwissen bezogen auf den Inhalt der Aufgaben eine wichtige Rolle (Perfetti 2007). Es ist davon auszugehen, dass die verschiedenen Teilaspekte individueller sprachlicher Fähigkeit (z. B. Wortschatz oder Strategiewissen) gemeinsam zum Einsatz kommen (vgl. Artelt et al. 2007).

Sprachlich rezeptive Fähigkeiten sind für die Bearbeitung der Aufgaben zur Evaluation der Bildungsstandards relevant, da wesentliche Teile der zur Lösung notwendigen Fachinformationen vorgegeben sind (vgl. Abb. 1). Diese müssen von den Probandinnen und Probanden zunächst angemessen verarbeitet werden, um tatsächlich hilfreich zu sein, da die Aufgaben letztlich als – wenn auch sehr kurze – Texte angesehen werden können. Darüber hinaus sind aber auch sprachlich produktive Fähigkeiten zumindest dann vonnöten, wenn die Testpersonen schriftsprachliche Antworten bei offenen Antwortformaten generieren sollen. Dies zeigt sich unter anderem bei vergleichenden Analysen von offenen und geschlossenen Antwortfor-



Fachinformation

Körper ändern bei Abkühlung oder Erwärmung ihre Länge. Diese Veränderung ist von der Temperaturänderung und von dem Material abhängig, aus dem der Körper besteht. In der nachfolgenden Tabelle sind für einige Metalle Längenänderungen angegeben.

Längenänderung eines Körpers von 10 m Länge bei einer Temperaturänderung von 10 °C:

Stoff	Längenänderung
Eisen	1,2 mm
Zink	3,0 mm
Stahl	1,2 mm
Kupfer	1,7 mm

In einer Werkstatt wird bei 20 °C die Länge eines Kupferrohres gemessen und das Rohr zugeschnitten. Das Rohr wird als Warmwasserleitung verlegt. Um die Leitung zu isolieren, wird die Länge im Betriebszustand erneut mit demselben Stahlmessband gemessen. Dabei hat das Kupferrohr die Temperatur des Wassers von 60 °C. Das Messband hat immer noch die Temperatur von 20 °C.

Was kann man über den Wert der gemessenen Länge im Betriebszustand im Vergleich zur ersten Längenmessung sagen?

Kreuze an.

☐ Der Messwert ist kleiner.

☐ Der Messwert ist gleich.

☐ Der Messwert ist größer.

☐ Man kann keine Aussage darüber treffen.

Abb. 1 Beispielaufgabe zur Erfassung der Kompetenz *Umgang mit Fachwissen Physik*

maten (Härtig 2014a, b). Beide Aspekte, der Einfluss sprachlich rezeptiver Fähigkeiten auf Kompetenztests sowie der Einfluss des Antwortformats sollen zunächst theoretisch fundiert werden.

2.1 Einfluss sprachlicher Fähigkeiten auf Kompetenztests

Die PISA 2003 Studie bietet die Möglichkeit, zwischen sprachlichen Fähigkeiten (gemessen als Lesekompetenz), mathematischen und naturwissenschaftlichen Kompetenzen Beziehungen herzustellen. Sowohl in der internationalen Stichprobe als auch in der nationalen Gruppe für Deutschland zeigten sich hohe latente Korrelationen ($r = 0{,}63$ bis $0{,}87$) zwischen den jeweiligen Konstrukten (Leutner et al. 2004). Diese hohen latenten Korrelationen bleiben auch bestehen, wenn kognitive Grundfähigkeiten als gemeinsamer Einflussfaktor auspartialisiert werden, wobei naturwissenschaftliche Kompetenz auch in einer nichtmetrischen multidimensionalen Skalierung in der Nähe der Lesekompetenz verortet werden kann (Leutner et al. 2004).

Diese Ergebnisse legen nahe, dass die Performanz in naturwissenschaftlichen Tests von sprachlichen Fähigkeiten beeinflusst werden könnte, wie es für das Fach Chemie gezeigt wurde (Ropohl et al. 2015). Dabei fallen die latenten Korrelationen zwischen sprachlich rezeptiven Fähigkeiten und naturwissenschaftlicher Kompetenz zwar niedriger aus ($r = 0{,}54$ bis $0{,}66$). Eine höhere Korrelation zeigt sich aber insbesondere bei Aufgaben, bei denen zur Lösung relevante Informationen Teil des Itemstamms sind. Dieser Einfluss sprachlich rezeptiver Fähigkeiten lässt sich insbesondere vor dem Hintergrund der Textverständnisforschung interpretieren: Hintergrund hierfür ist die Anforderung, aus dem Text (d. h. dem Itemstamm) eine angemessene mentale Repräsentation zu bilden und diese an das Vorwissen anzuknüpfen (vgl. Schnotz 2006). Erst dann wäre eine Bearbeitung des Items möglich.

In verschiedenen Studien wurden die einzelnen Personenmerkmale (u. a. fachspezifisches Vorwissen und Facetten sprachlicher Fähigkeiten) zur Erklärung des Textverständnisses expositorischer Texte untersucht. Dabei fanden zum Beispiel Schaffner und Schiefele (2013) sowie Cromley et al. (2010) Evidenz für die Bedeutsamkeit von schlussfolgerndem Denken (bei einer erheblichen Differenz der verwendeten Maße), Strategiewissen und sprachlich rezeptiver Fähigkeiten, beziehungsweise Wortschatz und Worterkennung. Unterschiedlich fällt hingegen der Einfluss des fachspezifischen Vorwissens aus: Während sich in einem Strukturgleichungsmodell mit nur direkten Effekten bei Schaffner und Schiefele (2013) kein Einfluss zeigte, fanden Cromley et al. (2010) sowohl direkte wie mediierende Effekte. Ferner zeigten O'Reilly und McNamara (2007) einen Kompensationseffekt von sprachlichen Fähigkeiten auf das Vorwissen. Dabei waren Schülerinnen und Schüler in der Lage beim Lesen eines naturwissenschaftlichen Textes niedriges Vorwissen partiell mit hohen sprachlichen Fähigkeiten auszugleichen.

Mit Blick auf Schulleistungsstudien, wie auch die Evaluation der Bildungsstandards, sind große Teile dieser Befunde als relevant anzusehen. Die Testitems umfassen einen Aufgabenstamm, der Sachinformationen und meist eine kontextuelle Einbettung umfasst sowie die Handlungsanweisung (s. Abb. 1 für ein Beispielitem). Zudem müssen entweder Distraktoren gelesen oder aber schriftliche Antworten gegeben werden.

2.2 Einfluss des Antwortformats auf die Performanz

Im Rahmen der Evaluation der Bildungsstandards werden Aufgabentypen mit geschlossenen und offenen Antwortformaten eingesetzt, die unterschiedliche Anforderungen an die Testteilnehmenden stellen. Zur Bearbeitung von Multiple Choice Aufgaben können beispielsweise Strategien zum intelligenten Ausschließen von Antwortalternativen zum Tragen kommen (vgl. Klieme et al. 2000). Dabei müssen die Informationen des Aufgabenstamms mit den Antwortoptionen in Beziehung gesetzt werden, weshalb auch die Fähigkeit zum schlussfolgernden Denken vonnöten ist. Bei offenen Antwortformaten mit kurzer oder ausführlicher Antwort muss die Lösung eigenständig generiert und unter Umständen auch schriftlich begründet oder dargelegt werden, wobei dann sprachlich produktive Fähigkeiten relevant werden.

Generell führen offene Antwortformate zu niedrigeren Personenfähigkeiten als Multiple-Choice Items (DeMars 2000; Klieme et al. 2000; Leucht et al. 2012). Die Höhe dieses Unterschieds variiert allerdings auch aufgrund der nicht immer gegebenen Ähnlichkeit der Itemstämme (Rodriguez 2003). Härtig (2014a, b) verglich systematisch Physikaufgaben im Multiple-Choice Format mit Aufgaben im offenen Format. Dabei legte er Probandinnen und Probanden beide Varianten vor. Während sich bezogen auf das fachdidaktische Wissen Studierender für das Lehramt Physik nur sehr bedingt Effekte zeigten (Härtig 2014a), ließ sich der Einfluss des Antwortformats aber insbesondere bei einem reinen Fachtest bei Studierenden aller Fächer belegen, wobei die Testpersonen im Mittel in offenen Formaten schlechter abschnitten als in geschlossenen Formaten (Härtig 2014b).

Zusammenfassend kann festgehalten werden, dass Items mit geschlossenem Format eher Anforderungen an das Erkennen der korrekten Antwort stellen und solche mit offenem Format Anforderungen an das Generieren und Aufschreiben. Daher wird vermutet, dass sich bei allen Items rezeptive *sprachliche Fähigkeiten* auswirken, während offene Antwortformate zusätzlich produktive *sprachliche Fähigkeiten* erfordern. Insofern stellt sich die Frage, inwieweit das Antwortformat zusätzlich Auswirkungen auf die Performanz hat und wie diese im Zusammenhang mit sprachlichen Fähigkeiten stehen.

2.3 Ableitung der Forschungsfragen

Ausgehend von den Bildungsstandards wurde für deren Evaluation ein Kompetenzstrukturmodell entwickelt. Dieses beschreibt vier Kompetenzbereiche: *Umgang mit Fachwissen, Erkenntnisgewinnung, Kommunikation* und *Bewertung* (KMK 2005). Die Konzeption der Aufgaben zur Evaluation der Bildungsstandards sieht vor, zur Lösung relevante Fachinformationen zumindest in großen Teilen als Textfeld in den Itemstämmen vorzugeben (Kauertz et al. 2010; Sumfleth et al. 2013) (vgl. Abb. 1). Dieses bis dahin eher ungewöhnliche Itemformat führte innerhalb der Fachdidaktik zu kontroversen Diskussionen, ob die Aufgaben nicht zu stark *sprachliche Fähigkeiten* (mit) erfassen (Labudde et al. 2009). Exemplarisch wird dies hier für den Kompetenzbereich *Umgang mit Fachwissen* untersucht, da dieser sich einerseits von der Fähigkeit über Fachwissen zu verfügen, abgrenzt, andererseits aber darauf auch bezieht und daher gleichzeitig der Einfluss sprachlicher Fähigkeiten untersucht werden kann.

Da in den Aufgabenstämmen relevante Fachinformationen dargeboten werden, ist davon auszugehen, dass neben dem zugrunde liegenden Fachwissen sprachlich rezeptive Fähigkeiten die Performanz beeinflussen können. Begründen lässt sich dies unter anderem vor dem Hintergrund der Forschung zum Textverstehen. Daraus ergibt sich folgende Forschungsfrage:

(F.1) Inwieweit lässt sich die Leistung in den Items zur Messung des Umgangs mit Fachwissen Physik über Effekte des Fachwissens Physik durch sprachliche Fähigkeiten erklären?

Es ist nicht zwingend davon auszugehen, dass sich *sprachliche Fähigkeiten* gleichermaßen auf alle Aufgaben und bei allen Personen auswirken. Kauertz et al. (2010) vermuten, dass Personen mit niedrigen sprachlichen Fähigkeiten im Rahmen der Evaluation der Bildungsstandards benachteiligt werden. Dabei scheint sich das Antwortformat eines Items als kritische Größe herauszukristallisieren. So zeigte sich in früheren Arbeiten, dass bei Multiple-Choice Items andere konstruktirrelevante Merkmale von Testpersonen von Bedeutung sind als bei Items mit offenem Antwortformat (DeMars 2000; Rodriguez 2003). Bislang wurden in diesem Zusammenhang jedoch eher Merkmale aus den Bereichen Wissen oder Kompetenz untersucht – *sprachliche Fähigkeiten* wurden nicht betrachtet. Daraus ergeben sich zwei weitere Forschungsfragen:

(F.2) Zeigen sich bei Items zur Erfassung des Umgangs mit Fachwissen Physik Unterschiede in den Itemschwierigkeiten offener und geschlossener Antwortformate?

(F.3) Zeigen sich abhängig vom Aufgabenformat Unterschiede in der Relevanz sprachlicher Fähigkeiten für die Lösung von Items zur Erfassung des Umgangs mit Fachwissen Physik?

3 Methode

3.1 Stichprobe

Die Stichprobe mit $N = 1961$ Auszubildenden (23,3 % weiblich; Alter: $M = 18{,}40$, $SD = 2{,}31$) aus Berufen, in denen mathematisch-naturwissenschaftlichen Kompetenzen eine wichtige Rolle spielen (KFZ-Mechatroniker: 27,1 %, Elektroniker: 20,2 %, Industriemechaniker: 19,4 % und Industriekaufleute: 33,3 %), stammt aus der ersten Welle des Projekts ManKobE („*Ma*thematisch-*n*aturwissenschaftliche *Ko*mpetenzen in der *b*eruflichen *E*rstausbildung“, vgl. Retelsdorf et al. 2013). Die Daten wurden Ende 2012/Anfang 2013 an 64 beruflichen Schulen in den Bundesländern Baden-Württemberg, Bayern und Hessen erhoben.

3.2 Beschreibung der Instrumente

Zur Untersuchung der Forschungsfragen wurden drei Instrumente benötigt, wodurch eines als Gegenstand der Untersuchung bereits festgelegt war: Ein Kompetenztest für den Kompetenzbereich *Umgang mit Fachwissen Physik* (*UFP*), ferner ein Instrument

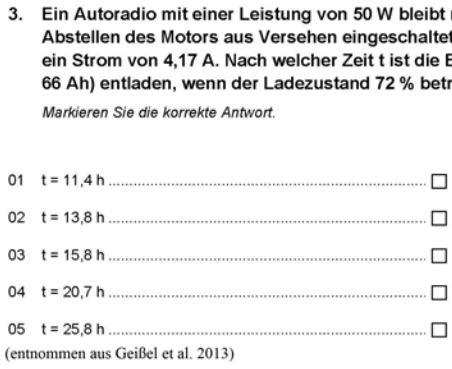
zur Erfassung des *Fachwissen Physik* (*FP*) sowie ein Instrument zur Erfassung der *sprachlichen Fähigkeiten* (*SF*).

UFP. Es konnte eine Auswahl von 40 Items der Testitems zur Evaluation der Bildungsstandards im Fach Physik genutzt werden, die in drei Aufgabenblöcken gebündelt waren. Die Testaufgaben wurden vom Institut zur Qualitätsentwicklung im Bildungswesen zur Verfügung gestellt. Alle Blöcke waren in Bezug auf verschiedene Itemcharakteristika wie das Antwortformat oder die Kompetenzteilbereiche gleichverteilt und repräsentativ für den Aufgabenpool der Bildungsstandards. Davon sind 27 Items im geschlossenen oder Kurzantwortformat (z. B. eine Zahl als Rechenergebnis oder ein einzelnes Wort), 13 Items erfordern längere schriftliche Antworten. Ein Beispiel ist in Abb. 1 dargestellt. Die EAP-Reliabilität betrug 0,71.

Blöcke mit insgesamt 25 Items aus den Zwischen- beziehungsweise Abschlussprüfungen der Industrie- und Handelskammer (IHK) für technische Berufe ausgewählt, die ausschließlich mit den Physikunterrichtsinhalten der Sekundarstufe I lösbar sein sollten. Dabei wurden solche Aufgaben ausgewählt, die einerseits inhaltlich für die Stichprobe relevant sind (Mechanik und Elektrikzitätslehre) und andererseits auch für die Inhalte des Physikunterrichts der Sekundarstufe I repräsentativ. Im Gegensatz zu den *UFP* Items ist hier kein Vorwissen vorgegeben. Das zur Lösung benötigte Wissen (z. B. Formeln oder konzeptuelles Verständnis) müssen die Probandinnen und Probanden mitbringen, ansonsten ist eine Lösung eher unwahrscheinlich. Ein Beispiel findet sich in Abb. 2. Die EAP-Reliabilität betrug 0,59.[2]

SF. Jude et al. (2008) unterteilen *sprachliche Fähigkeiten* nach Prozessen und Modalitäten. Wie oben ausgeführt sind für die im Artikel behandelten Fragestellungen nur rezeptive und produktive Anteile schriftsprachlicher Fähigkeiten relevant. Es ist nicht Ziel der Untersuchung zwischen einzelnen Anteilen der sprachlichen Fähigkeiten zu differenzieren, vielmehr soll hier erste Evidenz für den prinzipiellen Einfluss sprachlicher Fähigkeiten gesammelt werden. Es wird daher ein Instrument benötigt, das möglichst breit und ökonomisch beide Prozesse sprachlicher Fähigkeiten erfasst.

Dazu wurden drei C-Tests eingesetzt, die allgemein als zuverlässige und praktikable Instrumente zur Erfassung allgemeiner Sprachfähigkeit gelten (z. B. Raatz und Klein-Braley 2002; Eckes und Grotjahn 2006). So kommt auch Asano (2014) zu der Schlussfolgerung, dass „C-Tests nicht nur die Lesekompetenz oder die Schreibkompetenz messen, sondern vornehmlich integrative Fähigkeiten“ (S. 50). C-Tests


3. **Ein Autoradio mit einer Leistung von 50 W bleibt nach dem Abstellen des Motors aus Versehen eingeschaltet. Dabei fließt ein Strom von 4,17 A. Nach welcher Zeit t ist die Batterie (12 V / 66 Ah) entladen, wenn der Ladezustand 72 % beträgt?**

Markieren Sie die korrekte Antwort.

01 t = 11,4 h ☐

02 t = 13,8 h ☐

03 t = 15,8 h ☐

04 t = 20,7 h ☐

05 t = 25,8 h ☐

(entnommen aus Geißel et al. 2013)

Abb. 2 Beispielaufgabe zur Erfassung des physikbezogenen Wissens

bestehen aus mehreren kurzen Lückentexten unterschiedlicher Thematik, in denen in mehreren Wörtern Teile des Wortes getilgt und sinngemäß rekonstruiert werden müssen. Sie können somit rezeptive und produktive Aspekte sprachlicher Fähigkeiten erfassen.

Bei der Auswertung von C-Tests, kann dabei entweder nur die inhaltliche Richtigkeit bewertet werden, oder es werden zusätzlich orthografische oder grammatikalische Fehler berücksichtigt. Da für die drei Fragestellungen sowohl rezeptive (Lesen von Itemstämmen und Items) als auch produktive (Produktion offener Antworten) Fähigkeiten relevant sind, wurden die Antworten der Auszubildenden unter Berücksichtigung orthografischer und grammatikalischer Fehler ausgewertet. Für diese Form der Auswertung werden zudem reliablere Testscores postuliert (vgl. Eckes und Grotjahn 2006).[3] In der vorliegenden Studie wurden Tests von Wockenfuß und Raatz (2006) eingesetzt, die insgesamt 58 Items umfassen. Die EAP-Reliabilität betrug 0,85.

4 Statistische Analysen

Zur Beantwortung der Fragestellungen wurde eine Folge von drei zweiparametrischen logistischen Modellen geschätzt, die sich in ihren Ladungen der Items zur Erfassung des Kompetenzbereichs *UFP* unterscheiden (s. Abb. 3). Für alle latenten Faktoren gilt $\theta = 0$ und $\psi = 1$. Alle Items waren dichotom (richtig/falsch) kodiert. Die Analysen wurden in M*plus* unter Verwendung eines robusten Maximum-Likelihood-Schätzers (MLR, Yuan und Bentler 2000) durchgeführt. Die Tests zum *FP* und zum *UFP* wurden in einem rotierten Testheftdesign administriert, bei dem jede Person je einen Testblock aus jedem Bereich zu bearbeiten hatte. Für beide Bereiche gab es dabei je drei Blöcke (*UFP* mit 13 bzw. 14 Items pro Block, *FP* mit je neun Items pro Block, wobei zwei Items im Zuge der Skalierungen für ManKobE aufgrund schlechter Kennwerte eliminiert wurden). In Modell 1 wurden die drei latenten Variablen Kompetenzbereich *UFP, FP* und *SF*, so spezifiziert, dass nur Ladungen der jeweils zugehörigen Items auf eine latente Variable zugelassen wurden. Alle möglichen Doppelladungen wurden auf Null fixiert. In Modell 2 wurden zusätzlich Ladungen aller *UFP*-Items auf den *FP*-Faktor zugelassen. In Modell 3 werden schließlich auch die Ladungen aller *UFP*-Items auf den *SF*-Faktor zugelassen. Die konkurrierenden Modelle wurden mittels Loglikelihood-Ratio-Test verglichen.

Zu beachten ist, dass die einzelnen Items in C-Tests nicht notwendig lokal stochastisch unabhängig sind (Harsch und Hartig 2010; Schroeders et al. 2014). Allerdings scheinen diese Abhängigkeiten zumeist eher klein zu sein und sich in erster Linie auf Itemschwierigkeiten und Reliabilitäten auszuwirken (Schroeders et al. 2014).[4]

5 Ergebnisse

5.1 Fragestellung 1

Im Folgenden werden die Vergleiche der drei oben genannten Modelle mit zusätzlichem Blick auf die Ladungsmuster der Items zur Erfassung des Kompetenzbereichs

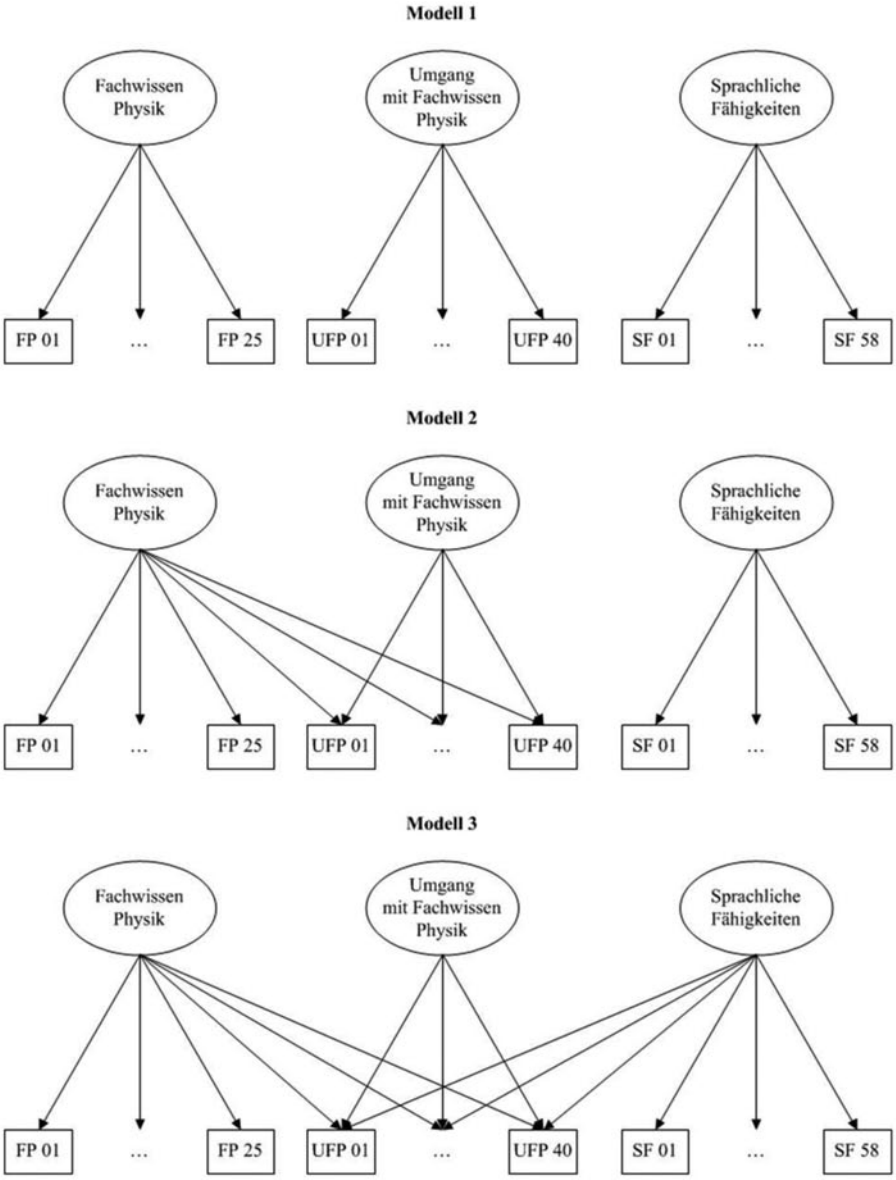


Abb. 3 Sequenz der geschätzten faktorenanalytischen Modelle

UFP beschrieben. In Tab. 1 sind die Ergebnisse der Modellvergleiche dargestellt. Zunächst zeigte sich, dass in Modell 1, dem Modell ohne Doppelladungen, alle Items zur Erfassung des Kompetenzbereichs *UFP* wie erwartet positiv, signifikant und substanziell auf den zugehörigen Faktor laden. Die Korrelationen zwischen den drei Fak-

Tab. 1 Vergleich der Modellanpassungen von Modell 1 bis 3

Modelle	Loglikelihood	Anzahl freier Parameter	Korrekturfaktor	$-2LR_{korrigiert}$	*p*
Modell 1	−64464,533	249	1,026	–	–
Modell 2	−64375,231	288	1,055	143,699	0,000
Modell 3	−64331,280	327	1,047	89,262	0,000

Die Modellvergleiche wurden sukzessive durchgeführt, d. h. Modell 2 wurde mit Modell 1 und Modell 3 mit Modell 2 verglichen

toren betrugen $r(UFP, FP) = 0{,}82$, $r(UFP, SF) = 0{,}58$ und $r(FP, SF) = 0{,}44$. In Modell 2 wurden zusätzlich Doppelladungen der Items *UFP* auf den Faktor *FP* zugelassen. Dieses Modell passte signifikant besser auf die Daten als Modell 1. Zudem zeigte sich, dass der *FP*-Faktor erhebliche Anteile der Varianz in den Items zur Erfassung des Kompetenzbereichs *UFP* aufklärte. So luden von den 40 Items zur Erfassung des Kompetenzbereichs 37 signifikant und positiv auf den *FP*-Faktor, bei zwei weiteren Items wurden die Ladungen knapp nicht signifikant ($p = 0{,}066$ bzw. $p = 0{,}073$). Lediglich ein Item lud nicht signifikant auf den *FP*-Faktor ($p = 0{,}326$). Gleichzeitig zeigten sich nur noch für 6 der 40 Items signifikante Ladungen auf den *UFP*-Faktor, ein weiteres Item verpasste das Signifikanzniveau knapp ($p = 0{,}064$). Der Großteil der Items zeigte in Modell 2 keine signifikanten Ladungen mehr auf den *UFP*-Faktor ($p \geq 0{,}101$). Modell 3 schließlich passte wiederum besser auf die Daten als Modell 2. Hier zeigten sich bis auf vier Items weiterhin durchgängig signifikante positive Ladungen der Items zur Erfassung des Kompetenzbereichs *UFP* auf den Fachwissensfaktor *FP*. Von den vier nichtsignifikanten Items wurde eine Ladung knapp nicht signifikant ($p = 0{,}062$), während die verbleibenden drei Items das Signifikanzniveau deutlich verpassten ($p \geq 0{,}115$). Zudem ergaben sich bei 12 Items signifikante Ladungen auf den Faktor *SF*, für vier weitere Items zeigten sich immerhin noch Tendenzen ($p \leq 0{,}090$). Für diese zwölf bis 16 Items zeigte sich also, dass unter Kontrolle von *FP* *SF* zusätzliche Varianz aufklären. Merkmale dieser Items werden im Folgenden zur Beantwortung der beiden verbleibenden Fragestellungen diskutiert.[5]

5.2 Fragestellung 2

Die Analysen zu Unterschieden in Itemschwierigkeiten offener und geschlossener Antwortformate zeigten, dass die Ladungen der Items mit offenem Antwortformat auf den Faktor *SF* höher ausfallen als die der übrigen Items (s. Abb. 4a). Zudem zeigte sich, dass die mittlere unstandardisierte Ladung aller 13 Items mit offenem Antwortformat ($M(\lambda) = 0{,}319$, $SE = 0{,}091$) signifikant höher ausfiel als die mittlere unstandardisierte Ladung der 27 übrigen Items ($M(\lambda) = 0{,}139$, $SE = 0{,}086$; Wald Chi²-Test: $\chi^2(1) = 5{,}858$, $p < 0{,}05$). Noch deutlicher fiel der Unterschied mit Blick auf die Lösungswahrscheinlichkeiten der Items zur Erfassung des Kompetenzbereichs *UFP* aus (siehe Abb. 4b). Die mittlere Lösungswahrscheinlichkeit für Items mit offenen Antwortformaten betrug $P = 0{,}388$ ($SE = 0{,}008$), für die übrigen Items $P = 0{,}631$ ($SE = 0{,}006$), Wald Chi²-Test: $\chi^2(1) = 959{,}893$, $p < 0{,}001$.

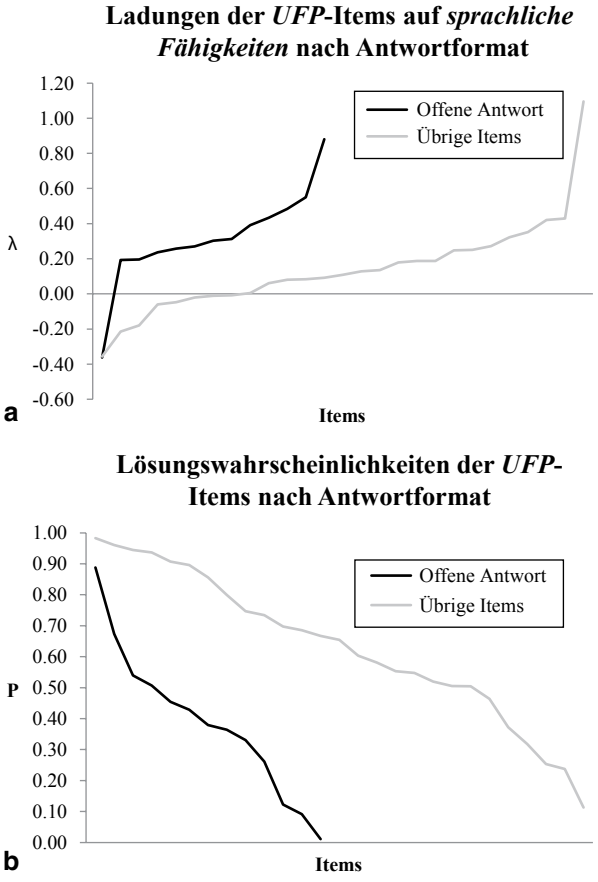


Abb. 4 Items zur Erfassung des Kompetenzbereichs *Umgang mit Fachwissen Physik* (*UFP*) nach Antwortformat (Anmerkung: **a** Faktorladungen auf den Faktor *sprachliche Fähigkeiten* und **b** Lösungswahrscheinlichkeiten bei mittleren Fähigkeiten in allen drei Dimensionen. Die Parameter sind in beiden Abbildungen der Größe nach sortiert.)

5.3 Fragestellung 3

Zur Beantwortung von Fragestellung drei wurden die mittleren Item Characteristic Curves der Items zur Erfassung des Kompetenzbereichs *UFP* in Abhängigkeit von den sprachlichen Fähigkeiten nach Antwortformat gebildet (s. Abb. 5). Es wird deutlich, dass die Lösungswahrscheinlichkeit für beide Antwortformate in Abhängigkeit von den *SF* stieg, für die offenen Antworten jedoch in deutlichem höherem Ausmaß.

6 Diskussion

Die Bildungsstandards für den mittleren Schulabschluss für die naturwissenschaftlichen Fächer beschreiben Kompetenzen. Aktuell werden Aufgaben eingesetzt, um die Bildungsstandards beispielsweise zum Kompetenzbereich *Umgang mit Fachwissen* zu evaluieren. Sowohl mit Rücksicht auf die unterschiedlichen Lehrpläne, als auch als Merkmal der Kompetenzorientierung beinhalten alle Aufgaben zur Evaluation dieses Kompetenzbereichs für die Lösung relevante Fachinformationen (vgl. Abb. 1)

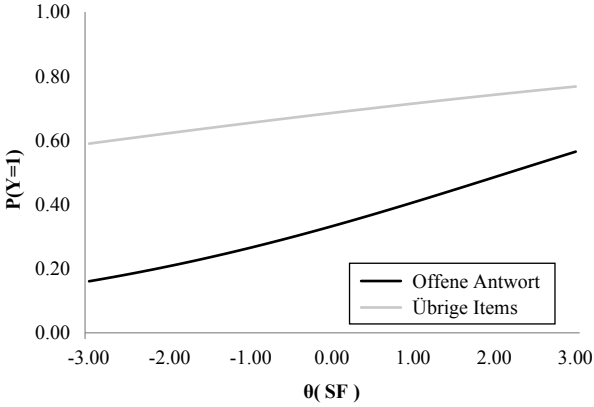


Abb. 5 Mittlere Item Characteristic Curves der Items zur Erfassung des Kompetenzbereichs *Umgang mit Fachwissen Physik* in Abhängigkeit der *sprachlichen Fähigkeiten* (*SF*)

(Sumfleth et al. 2013). Durch dieses Aufgabendesign lässt sich eine Überlagerung mit schriftsprachlich rezeptiven Fähigkeiten vermuten. Darüber hinaus sind zumindest für die Aufgaben mit offenem Antwortformat gegebenenfalls auch schriftsprachlich produktive Fähigkeiten relevant. In der vorliegenden Studie wurde diese Überlagerung im Rahmen von drei Forschungsfragen näher untersucht. Dabei ist festzuhalten, dass die hier verwendete Stichprobe nicht repräsentativ für alle Schülerinnen und Schüler mit mittlerem Bildungsabschluss ist, da es sich um Auszubildende zweier Berufsfelder handelte. Zwar deutete deren Performanz in den Aufgaben zur Evaluation der Bildungsstandards eine gewisse Vergleichbarkeit an, dies müsste jedoch in Folgeuntersuchungen validiert werden.

Zunächst zeigte sich Evidenz dafür, dass neben dem zugrunde liegenden *Fachwissen Physik* im Sinne eines inhaltsspezifischen Vorwissens tatsächlich für einen relevanten Teil der Bildungsstandards Items auch *sprachliche Fähigkeiten* relevant sind. Dies steht im Einklang zu einer Untersuchung von Ropohl et al. (2015) im Fach Chemie. Es lässt sich schlussfolgern, dass für diese Items sprachlich rezeptive Fähigkeiten relevant sind, teilweise auch sprachlich produktive Fähigkeiten. Ersteres kann dann der Fall sein, wenn die zur Verfügung gestellten Fachinformationen im Itemstamm tatsächlich zur Bearbeitung des Items herangezogen werden, da hier zunächst eine mentale Repräsentation in Einklang mit dem Vorwissen zu bringen ist. Letzteres ist der Fall, wenn das Item im offenen Antwortformat gestellt ist. Dieser Teilaspekt wurde in der zweiten Forschungsfrage näher untersucht. Tatsächlich zeigte sich im Einklang mit ähnlichen Untersuchungen (DeMars 2000; Rodrigues 2003) ein Einfluss des Antwortformats: Items im offenen Format sind im Mittel signifikant schwieriger als Items im geschlossenen Antwortformat. Dabei ist zu berücksichtigen, dass Items mit offenem Antwortformat teilweise höhere Anforderungen an die Testpersonen stellen. Dies ist der Fall, wenn sie eingesetzt werden, um Items mit einer hohen Komplexität zu testen für die keine geeigneten Distraktoren im Rahmen eines Multiple Choice Formats entwickelt werden können. Es zeigte sich aber auch, dass bei Items im offenen Antwortformat eher Varianz durch *sprachliche Fähigkeiten* erklärt wird als den geschlossenen Items, was dagegen spricht, das ausschließlich die höhere Komplexität des Fachinhalts die Performanz beeinflusst. Ferner konnte im Rahmen der dritten Forschungsfrage ein Indiz für einen Interaktionseffekt gefunden werden.

Generell sank die Aufgabenschwierigkeit mit steigenden sprachlichen Fähigkeiten, insbesondere für die Items im offenen Antwortformat.

Zusammenfassend finden wir Evidenz dafür, dass die Items zur Evaluation der Bildungsstandards zumindest im Kompetenzbereich *Umgang mit Fachwissen Physik* nicht nur physikbezogenes Fachwissen erfassen, sondern auch in relevantem Umfang *sprachliche Fähigkeiten*. Dieses Ergebnis fügt sich in den internationalen, fachdidaktischen Diskurs ein, bei dem *sprachliche Fähigkeiten* (oftmals im Sinne eines Textverständnisses) als zentral für den Erwerb naturwissenschaftlicher Kompetenzen angesehen werden: „Language is an essential technology and thus an integral part of science and science literacy, particularly written language" (Yore et al. 2004, S. 348). Schließlich konstituieren sich die Naturwissenschaften in Texten und ein Verständnis für Texte sowie geeignete Strategien zum Umgang damit sind essentiell für die Entwicklung naturwissenschaftlichen Wissens (vgl. Norris und Phillips 2003).

Des Weiteren sind die Ergebnisse vor dem Hintergrund der aktuell debattierten Sprachförderung im Schulunterricht bedeutsam, die im Zusammenhang mit fachlichen Leistungen stehen. Beispielsweise zeigte sich für den „IQB-Ländervergleich 2012" ein bedeutender Effekt für Schülerinnen und Schüler mit Zuwanderungshintergrund mit der Leistung im Fach Mathematik und den naturwissenschaftlichen Fächern (Pöhlmann et al. 2013). Lernende mit Zuwanderungshintergrund schnitten insbesondere dann schlechter ab, wenn im Elternhaus überwiegend die Muttersprache der Eltern und nicht Deutsch gesprochen wird. Ferner zeigt sich auch hier in Regressionsanalysen zumindest hypothetisch der Einfluss sprachlicher Fähigkeiten: Kontrolliert man die Sprache im Elternhaus, sinken die Disparitäten für die Kinder mit Zuwanderungshintergrund. Hierbei wird allerdings angenommen, dass es einen direkten Zusammenhang zwischen Sprache im Elternhaus und sprachlichen Fähigkeiten in Deutsch gibt. Durch die Hinzunahme der sprachlichen Fähigkeiten werden auch in unserer Studie die Effekte zum Beispiel des Antwortformats nicht völlig erklärt.

In diesem Kontext scheint relevant, dass es sich um expositorische Texte handelt. Aktuell widmen sich mehrere Arbeiten der Frage, ob Sprachförderung nicht auch in den verschiedenen Unterrichtsfächern situiert werden muss, da sich die Textmerkmale von expositorischen und narrativen Texten mitunter erheblich unterscheiden (z. B. Agel et al. 2012). Diese Annahme könnte auch hier eine Erklärung geben. Sowohl die verwendeten Fachinformationen im Itemstamm sind physikbezogen expositorisch, als auch schriftsprachliche Antworten entsprechend zu formulieren. Wollte man dies gezielt berücksichtigen, wäre der verwendete C-Test nur eine Annäherung an die benötigten sprachlichen Fähigkeiten, da dort im Kern narrative Texte verwendet werden. Ferner muss einschränkend hinzugefügt werden, dass der C-Test produktive und rezeptive schriftsprachliche Fähigkeiten als gemeinsam erfasst. Eine Ausdifferenzierung in die einzelnen Teilprozesse, wie Worterkennung und schlussfolgerndem Denken wäre nun in Anlehnung an Arbeiten zum Textverständnis ein nächster notwendiger Schritt.

Insgesamt lässt sich feststellen, dass zur Evaluation der Bildungsstandards zumindest für das Fach Physik neben rein fachbezogenen Kompetenzen auch *sprachliche Fähigkeiten* in relevantem Umfang miterfasst werden. Hier ergeben sich nun verschiedene Anschlussfragestellungen, vor allem über die Funktion und Bedeu-

tung sprachlicher Fähigkeiten im Fachunterricht generell, die Gegenstand weiterer Forschung sein sollten (Härtig et al. 2015). Dabei ist auch zu klären, inwiefern die Fähigkeit, physikalisches Sachwissen aus einem Aufgabenstamm zu entnehmen eine didaktisch bedeutsame Fähigkeit darstellt oder nur für ein besseres Verständnis der Testung bedeutsam ist.

Anmerkungen

1 Der besseren Lesbarkeit halber wird im folgenden Text auf das Präfix „schrift“ meistens verzichtet, alle Untersuchungen und Ergebnisse beziehen sich aber ausschließlich auf die schriftliche Modalität.

2 Die relativ niedrigen Reliabilitäten der Tests zum Umgang mit Fachwissen Physik und Fachwissen Physik sind vermutlich auf das Rotationsdesign (s. Abschnitt zu statistischen Analysen) zurückzuführen und stellen eine Unterschätzung der Reliabilität dar.

3 Zusätzlich wurden die Analysen für die Auswertungen ohne Berücksichtigung orthografischer und grammatikalischer Fehler wiederholt. Die Ergebnisse sind cum grano salis vergleichbar und können bei Interesse von den Autoren/der Autorin angefordert werden.

4 Im Kontext dieser Untersuchung dürfte die mögliche Verletzung dieser Annahme eher nicht ins Gewicht fallen, da hier die dominante Dimension (vgl. Nandakumar 1991) sprachlicher Fähigkeiten von Interesse ist und nicht eine differenzierte Analyse der C-Tests.

5 Zusätzlich zu der relevanten Modellsequenz für die konkreten Fragestellungen wurden zwei weitere Modelle jeweils ausgehend von Modell 1 geschätzt, bei denen einmal ausschließlich für die UFP-Items (Modell 1a) und einmal ausschließlich für die FP-Items (Modell 1b) Doppelladungen auf den Faktor SF zugelassen wurden. In Modell 1a zeigten sich für alle 40 UFP-Items positive Doppelladungen, von denen 36 signifikant wurden. In Modell 1b zeigten sich für 11 von 25 Items keine signifikanten Doppelladungen; von den nichtsignifikanten Ladungen lagen zudem 5 im negativen Bereich.

Literatur

Agel, C., Beese, M., & Krämer, S. (2012). Naturwissenschaftliche Sprachförderung. *Der mathematische und naturwissenschaftliche Unterricht, 65*, 36–43.

Asano, Y. (2014). C-Tests und „allgemeine Sprachkompetenz“: Theoretische Überlegungen und empirische Analysen. In R. Grotjahn (Hrsg.), *Der C-Test: Aktuelle Tendenzen* (S. 41–54). Frankfurt a. M.: Lang.

Cromley, J. G. (2009). Reading achievement and science proficiency: International comparisons from the Programme on International Student Assessment. *Reading Psychology, 30*, 89–116.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 102*, 687–700.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55–77.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*, 290–325.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10*, 301–322.

Geißel, B., Nickolaus, R., Stefanica, F., Härtig, H., & Neumann, K. (2013). Die Relevanz mathematischer und naturwissenschaftlicher Kompetenzen für die fachliche Kompetenzentwicklung in gewerblich-technischen Berufen. In R. Nickolaus, J. Retelsdorf, E. Winther, & O. Köller (Hrsg.), *Mathematisch-naturwissenschaftliche Kompetenzen in der beruflichen Erstausbildung* (Zeitschrift für Berufs- und Wirtschaftspädagogik: Beiheft 26, S. 39–66). Stuttgart: Franz Steiner Verlag.

Hall, S. S., Kowalski, R., Paterson, K. B., Basran, J., Filik, R., & Maltby, J. (2015). Local text cohesion, reading ability and individual science aspirations: Key factors influencing comprehension in science classes. *British Educational Research Journal, 41*, 122–142.

Harsch, C., & Hartig, J. (2010). Empirische und inhaltliche Analyse lokaler Abhängigkeiten im C-Test. In R. Grotjahn (Hrsg.), *Der C-Test: Beiträge aus der aktuellen Forschung* (S. 193–204). Frankfurt a. M.: Lang.

Härtig, H. (2014a). Software-basierte Evaluation freier Antwortformate. *Zeitschrift für Didaktik der Naturwissenschaften, 20*, 115–128.

Härtig, H. (2014b). Der Force Concept Inventory Vergleich einer offenen und einer geschlossenen Version. *PhyDid A, 13*, 53–61.

Härtig, H., Bernholt, S., Prechtl, H., & Retelsdorf, J. (2015) Unterrichtssprache im Fachunterricht – Stand der Forschung und Forschungsperspektiven am Beispiel des Textverständnisses. *Zeitschrift für Didaktik der Naturwissenschaften*, 1–13. doi: 10.1007/s40573-015-0027-7.

Jude, N., Klieme, E., Eichler, W., Lehmann, R. H., Nold, G., & Schröder, K. (2008). Strukturen sprachlicher Kompetenzen. In E. Klieme (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (S. 191–201). Weinheim: Beltz.

Kauertz, A., Fischer, H. E., & Jansen, M. (2013). Kompetenzstufenmodelle für das Fach Physik. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 92–100). Münster: Waxmann.

Klieme, E., Baumert, J., Köller, O., & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos, & R. H. Lehmann (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (Bd. 1, S. 85–133). Opladen: Leske + Budrich Verlag.

KMK. (2003) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2003). *Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss* (Jahrgangsstufe 10). http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mittleren-SA.pdf. Zugegriffen: 02. Jan. 2010.

KMK. (2005)= *Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). Bildungsstandards im Fach Physik für den Mittleren Schulabschluss (Beschluss vom 16.12.2004)*. München: Luchterhand.

Labudde, P., Duit, R., Fickermann, D., Fischer, H., Harms, U., & Mikelskis, H. (2009). Schwerpunkttagung „Kompetenzmodelle und Bildungsstandards: Aufgaben für die naturwissenschaftsdidaktische Forschung". *Zeitschrift für Didaktik der Naturwissenschaften, 15*, 343–370.

Leucht, M., Harsch, C., Pant, H. A., & Köller, O. (2012). Steuerung zukünftiger Aufgabenentwicklung durch Vorhersage der Schwierigkeiten eines Tests für die erste Fremdsprache Englisch durch *Dutch Grid* Merkmale. *Diagnostica, 58*, 31–44.

Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen. In M. Prenzel, J. Baumert, W. Blum, R. H. 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. 146–175). Münster: Waxman.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99–117.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*, 224–240.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "highstakes" measures of high school students' science achievement. *American Educational Research Journal, 44*, 161–196.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357–383.

Pöhlmann, C., Haag, N., & Stanat, P. (2013). Zuwanderungsbezogene Disparitäten. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 297–329). Münster: Waxmann.

Raatz, U., & Klein-Braley, C. (2002). Introduction to language testing and to C-Tests. In J. A. Coleman, R. Grotjahn, & U. Raatz, (Hrsg.), *University language testing and the C-test* (S. 75–91). Bochum: AKS-Verlag.

Retelsdorf, J., Lindner, C., Nickolaus, R., Winther, E., & Köller, O. (2013). Forschungsdesiderate und Perspektiven – Ausblick auf ein Projekt zur Untersuchung mathematisch-naturwissenschaftlicher Kompetenzen in der beruflichen Erstausbildung (ManKobE). In R. Nickolaus, J. Retelsdorf, E. Winther, & O. Köller (Hrsg.), *Mathematisch-naturwissenschaftliche Kompetenzen in der beruflichen Erstausbildung* (Zeitschrift für Berufs- und Wirtschaftspädagogik: Beiheft 26, S. 227–234). Stuttgart: Franz Steiner Verlag.

Rodriguez, M. C. (2003). Construct equivalence of Multiple-Choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*, 163–184.

Ropohl, M. (2010). *Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion*. Berlin: Logos.

Ropohl, M., Walpuski, M., & Sumfleth, E. (2015). Welches Aufgabenformat ist das richtige? – Empirischer Vergleich zweier Aufgabenformate zur standardbasierten Kompetenzmessung. *Zeitschrift für Didaktik der Naturwissenschaften*. doi:10.1007/s40573-014-0020-6.

Schaffner, E., & Schiefele, U. (2013). The prediction of reading comprehension by cognitive and motivational factors: Does text accessibility during comprehension testing make a difference? *Learning and Individual Differencies, 26*, 42–54.

Schmeck, A. (2011). *Visualisieren naturwissenschaftlicher Sachverhalte: Der Einsatz von vorgegebenen und selbst generierten Visualisierungen als Textverstehenshilfen beim Lernen aus naturwissenschaftlichen Sachtexten*. Essen: Universität Duisburg-Essen. http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-25873/Diss_Schmeck.pdf. Zugegriffen: 15. April 2015.

Schnotz, W. (2006). Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverständnis aus der Sicht der Psychologie und der kognitiven Linguistik. In H. Blühdorn, E. Breindl, & U. H. Waßner (Hrsg.), *Text - Verstehen. Grammatik und darüber hinaus* (S. 222–238). Berlin: Walter de Gruyter.

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-tests. *Journal of Educational Measurement, 51*, 400–418.

Sumfleth, E., & Tiemann, R. (2000). Bilder und Begriffe – Repräsentieren sie ähnliche Inhalte? *Zeitschrift für Didaktik der Naturwissenschaften, 6*, 115–127.

Sumfleth, E., Klebba, N., Kauertz, A., Mayer, J., Fischer, H. E., & Walpuski, M. (2013). Beschreibung der untersuchten naturwissenschaftlichen Kompetenzen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 38–52). Münster: Waxmann.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

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

Wockenfuß, V., & Raatz, U. (2006). Über den Zusammenhang zwischen Testleistung und Klassenstufe bei muttersprachlichen C-Tests. In R. Grotjahn (Hrsg.), *Der C-Test: Theorie, Empirie, Anwendungen* (S. 211–242). Frankfurt a. M.: Lang.

Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J., Treagust, D. F., et al. (2004). New directions in language and science education research. *Reading Research Quarterly, 39*, 347–352.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodolgy, 30*, 167–202.