



Modellierung der Struktur der Variablenkontrollstrategie und Abbildung von Veränderungen in der Grundschule

Martina Brandenburger¹ · Cem Aydin Salim¹ · Martin Schwichow¹ · Jens Wilbers¹ · Silke Mikelskis-Seifert¹

Eingegangen: 15. April 2020 / Angenommen: 9. März 2022 / Online publiziert: 8. April 2022
© Der/die Autor(en) 2022

Zusammenfassung

Die Variablenkontrolle ist bei der Planung und Durchführung von Experimenten von besonderer Bedeutung, weil sie eindeutige Aussagen über Beziehungen zwischen Ursache und Wirkung zulässt. Ihre Anwendung ist daher ein eigenständiges Lernziel des naturwissenschaftlichen Sachunterrichts und Gegenstand zahlreicher empirischer Studien. Entsprechende Fähigkeiten werden unter dem Begriff Variablenkontrollstrategie (VKS) zusammengefasst und beinhalten die vier Teilfähigkeiten: 1) Planung kontrollierter Experimente, 2) Identifizierung kontrollierter Experimente, 3) Interpretation der Ergebnisse kontrollierter Experimente und 4) Verständnis der fehlenden Aussagekraft unkontrollierter Experimente. Bisherige Studien zeigen starke positive Veränderungen bezüglich der VKS während der Grundschulzeit. Allerdings erfassen sie oft nur eine Teilfähigkeit bzw. differenzieren in ihren Analysen nicht zwischen unterschiedlichen Teilfähigkeiten oder dem Einfluss der Fachkontexte der Aufgaben. Wir haben zur Erfassung der VKS in der Grundschule ein Testinstrument im Multiple-Choice-Format entwickelt, welches Aufgaben zu den Teilfähigkeiten Identifizierung und Interpretation in unterschiedlichen Fachkontexten enthält. Das Instrument wurde in einer Querschnittstudie mit $N=415$ Zweit- bis Viertklässler*innen eingesetzt. Entgegen bisherigen Befunden zeigen die Ergebnisse einer Rasch-Analyse eine mehrdimensionale Struktur der VKS entsprechend den Teilfähigkeiten. Die Fachkontexte der Aufgaben haben keinen Einfluss auf die Dimensionalität. Die Schwierigkeitsstruktur von Aufgaben wird durch die angesprochene Teilfähigkeit (Identifizierung ist einfacher als Interpretation) und den gewählten Aufgabentyp (z. B. Wahl der Distraktoren nach Schülervorstellungen) beeinflusst. Darüber hinaus wurde eine unterrichtliche Förderung der VKS untersucht ($N=44$), um abzuschätzen, inwiefern das entwickelte Testinstrument erwartete Veränderungen hinsichtlich der VKS abbildet. Die gemessenen Veränderungen werden in diesem Beitrag in Relation zur Querschnittstudie gesetzt. Abschließend werden die Konsequenzen unserer Befunde für die Messung und Förderung der VKS in der Grundschule diskutiert.

Schlüsselwörter Variablenkontrolle · Grundschule · Rasch-Analyse · Denk- und Arbeitsweisen

✉ Martina Brandenburger
martina.brandenburger@ph-freiburg.de

¹ Fachrichtung Physik, Pädagogische Hochschule Freiburg,
Kunzenweg 21, 79117 Freiburg, Deutschland

Modelling the Structure of the Control of Variables Strategy (CVS) and Mapping Changes in CVS through Elementary School

Abstract

Control variables is of particular relevance for the planning of experiments, because only controlled experiments allow valid statements about cause-effect relations. Skills associated with the design and interpretation of controlled experiments are summarized under the term control of variables strategy (CVS) and compass the four subskills: 1) planning controlled experiments, 2) identifying controlled experiments, 3) interpreting the results of controlled experiments and 4) understanding the invalidity of uncontrolled experiments. Existing studies evidence strong positive changes regarding CVS during elementary school. However, they only cover a single subskill or do not discriminate different subskills from the influence of task contexts on students CVS skills. Therefore, we developed a CVS test instrument for elementary school in multiple-choice format with tasks for the CVS subskills “identification” and “interpretation” in different contexts. The instrument was used in a cross-sectional study with $N=415$ second to fourth graders. Contrary to previous findings, the results of a Rasch analysis show a multidimensional structure of the CVS according to the two subskills. The task context has no influence on the dimensionality. The item difficulty depends on the subskills (identification is easier than interpretation) and the task type (e.g. choice of distractors according to misconceptions). Additionally we examined an instructional support ($N=44$) in order to assess the extent to which our test reflects expected changes in the VKS. The measured changes are set in relation to the development in our cross-sectional study. We discuss the consequences of our findings for the measurement and enhancement of CVS in elementary school.

Keywords Control of variables strategy · Elementary school · Rasch analysis · Inquiry skills

Einleitung

Ziel des Sachunterrichts in der Grundschule ist, neben der Weiterentwicklung von Präkonzepten in Richtung naturwissenschaftlicher Konzepte, die Förderung naturwissenschaftlicher Denk- und Arbeitsweisen (Bybee 1997; KMK 2015; GDSU 2013). Dabei kommt dem Experimentieren als zentrale Arbeitsweise der Naturwissenschaften eine besondere Bedeutung zu, weil Experimente, wenn sie kontrolliert sind, eindeutige Aussagen über Ursache-Wirkungs-Beziehungen ermöglichen (Kirchner 2013). Beim Experimentieren manipulieren Wissenschaftler*innen aktiv den Beobachtungsgegenstand, sodass sie zwei Bedingungen vergleichen können, die sich nur hinsichtlich einer Variablen unterscheiden. Wird nur eine Variable verändert – alle weiteren Variablen werden konstant gehalten – kann untersucht werden, ob die veränderte (unabhängige) Variable einen Einfluss auf eine abhängige Variable hat. Das Kontrollieren potenziell konfundierender Variablen (alternative Ursache-Wirkungs-Beziehungen) gewährleistet eine eindeutige Identifizierung kausaler Zusammenhänge und begründet die höhere Aussagekraft (Validität) von Experimenten im Vergleich zu reinen Beobachtungen (Woodward 2003). Entsprechend werden Experimente als Vergleich kontrollierter Bedingungen bevorzugt zur Prüfung von kausalen Hypothesen eingesetzt (Bohrmann 2017).

Um mit kontrollierten Experimenten kausale Hypothesen zu prüfen, sind vier Schritte notwendig: 1) Identifizierung derjenigen Variable, für die ein kausaler Effekt auf eine abhängige Variable untersucht werden soll, 2) Ver-

gleich von mindestens zwei Bedingungen, die sich in der Ausprägung dieser Variablen unterscheiden, 3) Sicherstellen, dass die Ausprägung aller anderen Variablen gleich sind und 4) Durchführung des Experiments und Beobachtung möglicher Unterschiede auf der abhängigen Variablen. Die beschriebenen Schritte werden unter dem Begriff Variablenkontrollstrategie (VKS) zusammengefasst (Siler und Klahr 2012).

Die Bedeutung der VKS sowie die Schritte hin zur Planung eines kontrollierten Experiments können wissenschaftstheoretisch begründet werden. Die Frage, inwiefern bereits Grundschul Kinder diese Schritte beherrschen bzw. erlernen können, ist hingegen empirisch zu untersuchen. Voraussetzung zur Klärung dieser Frage ist Wissen über die Struktur der VKS, d.h. Wissen über eine mögliche Unterteilung der VKS in trennbare Dimensionen. Eine solche Strukturierung fasst Fähigkeiten, die von Lernenden in ähnlicher Ausprägung beherrscht werden, zusammen und trennt diese von Fähigkeiten ab, die deutlich schwieriger oder einfacher sind. Auf diese Weise wird einerseits eine reduzierte Beschreibung von Fähigkeitsausprägungen durch die Zusammenfassung unterschiedlicher Fähigkeiten in eine Dimension ermöglicht. Andererseits wird die Struktur der VKS durch die Unterteilung in mehrere Teilfähigkeiten differenzierter abgebildet als durch ein eindimensionales Modell. Mithilfe einer solchen Strukturierung können ferner Fähigkeitsausprägungen und deren Veränderung im Laufe der Schulzeit bzw. in Folge von gezielten unterrichtlichen Interventionen beschrieben werden (Edelsbrunner und Dablander 2019). Ziel dieser Arbeit ist es, ein Struk-

turmodell für die VKS in der Grundschule vorzuschlagen und empirisch zu prüfen. Im Folgenden werden dazu zunächst empirische Befunde zur Ausprägung der VKS bei Grundschulkindern vorgestellt und mögliche Kategorien zur Strukturierung der VKS diskutiert.

Theoretischer Hintergrund

Empirische Befunde zur Ausprägung der VKS bei Grundschulkindern

Fragen nach abhängigen und unabhängigen Variablen sowie nach Variablenzusammenhängen spielen in der Unterrichtspraxis nur eine untergeordnete Rolle, da im naturwissenschaftlichen Unterricht oft ohne Hypothesen experimentiert wird (Nehring et al. 2016). Dementsprechend verwundert es nicht, dass Grundschul Kinder die VKS nicht konsistent anwenden können. Nichtsdestotrotz zeigten die Ergebnisse von Koerber et al. (2011) sowie Bullock (1991) positive Veränderungen bei der Anwendung der VKS bei Grundschulkindern zwischen der zweiten und vierten Jahrgangsstufe. Darüber hinaus führen unterrichtliche Interventionen, in welchen die VKS explizit adressiert und angewendet wird, bereits ab der zweiten Jahrgangsstufe zu einer deutlichen Steigerung entsprechender Fähigkeiten (Bohrmann 2017; Chen und Klahr 1999; Klahr und Nigam 2004).

Bei Betrachtung der Präkonzepte der Schüler*innen zur VKS fällt auf, dass bereits Grundschul Kinder ein intuitives Verständnis kontrollierter Experimente als „faire Vergleiche“ besitzen. Darüber hinaus ist vielen Grundschulkindern auch bekannt, dass Experimente genutzt werden, um Hypothesen bzw. Vermutungen zu überprüfen (für eine Übersicht siehe Zimmerman 2007). Auf der Grundlage dieser intuitiven Konzepte können Grundschul Kinder zwischen kontrollierten und unkontrollierten Experimenten zur Prüfung einer gegebenen Hypothese unterscheiden, wobei der Anteil an Schüler*innen, die eine solche Unterscheidung vornehmen können, zwischen der zweiten und vierten Jahrgangsstufe von 20 % auf mehr als 60 % steigt. Den Schüler*innen bereitet es allerdings trotzdem Probleme, kontrollierte Experimente selbständig zu planen (Bullock 1991; Bullock und Ziegler 1999; Koerber et al. 2011).

Zusammenfassend kann festgehalten werden, dass es Studien zur VKS in der Grundschule gibt, die positive Ergebnisse zur Entwicklung dieser Fähigkeit zeigen. Im Folgenden soll geklärt werden, was den Erfolg der Anwendung der VKS maßgeblich beeinflusst und ob eine Strukturierung empirisch abgebildet werden kann. Ferner stellt sich die Frage, ob es sich um eine domänenspezifische oder eine domänenübergreifende Fähigkeit handelt.

Variablenkontrolle als domänenspezifische Fähigkeit

Das Vorgehen bei der Planung kontrollierter Experimente zur Prüfung kausaler Hypothesen kann aus einer allgemeinen Kausalitätsdefinition abgeleitet werden. Demnach hat eine Variable dann einen kausalen Einfluss auf eine abhängige Variable, wenn ihre Veränderung eine Veränderung in der abhängigen Variablen erzeugt und andere Ursachen für diese Veränderung ausgeschlossen werden können (Woodward 2003). Da diese Definition von Kausalität nicht an einen spezifischen Fachkontext¹ gebunden ist, kann die VKS als eine domänenübergreifende Fähigkeit angesehen werden. Allerdings kann aus einer theoretischen Anwendbarkeit der VKS auf die Prüfung kausaler Hypothesen unabhängig von Fachkontexten nicht darauf geschlossen werden, dass dies Lernenden auch gelingt. So zeigten beispielsweise die Ergebnisse von Croker und Buchanan (2011) eine Abhängigkeit der VKS vom Vorwissen der Testpersonen bezüglich des Fachkontextes. Testpersonen wählten häufiger kontrollierte Experimente, wenn deren Erwartungen an den Ausgang des Experiments konsistent mit ihrem Vorwissen sind. Aus den Studien von Koslowski (1996) kann im Umkehrschluss geschlossen werden, dass Testpersonen bei der Interpretation von experimentellen Daten die VKS dann ignorieren, wenn die Ergebnisse ihren Erwartungen widersprechen. Domänenübergreifend bedeutet in diesem Zusammenhang also, dass eine Fähigkeit zwar grundsätzlich in unterschiedlichen Fachkontexten angewendet werden kann aber nicht, dass dies unabhängig vom Wissen über die Kontexte stattfindet (Wellnitz et al. 2017; Hetmanek et al. 2018; Zimmerman 2007).

Fachkontexte können durchaus einen entscheidenden Einfluss auf die Strukturierung von domänenübergreifenden Fähigkeiten haben. So zeigte sich beispielsweise bei der Evaluation der deutschen Bildungsstandards für den Kompetenzbereich „Naturwissenschaftliche Untersuchungen“ (Wellnitz et al. 2017), dass das Modell gut zu den Daten passt, welches die Aufgaben zu fachspezifischen Faktoren (Biologie, Chemie, Physik) und prozessbezogenen Fähigkeiten (Hypothesen generieren, Experimente planen, Experimente auswerten) zusammenfasst. Für die Schweizer Bildungsstandards kam Gut (2012) zu ähnlichen Ergebnissen. Demzufolge ist anzunehmen, dass auch die VKS als domänenspezifische Fähigkeit beschrieben werden kann und damit verbunden die Fachkontexte der Aufgaben bei der Prüfung eine Rolle spielen. Für den Grundschul-

¹ Ein „Fachkontext“ ergibt sich daraus, dass eine Aufgabe so in einen inhaltlichen Themenbereich eingebettet wird (z. B. Tiere beobachten), dass zur erfolgreichen Bearbeitung der Aufgabe der Rückgriff auf ein Konzept aus diesem Inhaltsbereich (z. B. Geschwindigkeit in der Physik) erforderlich ist.

Tab. 1 Übersicht über existierende VKS Testinstrumente für die Primarstufe

Test	VKS Teilfähigkeiten				Total	JG	Antwortformat	Fachkontext
	PL	ID	IN	VER				
„Wir machen Experimente“ (Edelsbrunner et al. 2018)	–	7	2	7	16	2–4	MC, MC+ Begründung	Physik, Alltag
Test zur Erfassung der VKS (Viefers et al. 2018)	12	12	4	8	23	3–4	MC, Auswertung von Laborheften zu geplanten und durchgeführten Experimenten	Physik
„Wir experimentieren“ (Bohrmann 2017)	3	1	3	5	12	3	MC, OA, MC+ Begründung	Physik, Alltag
Test experimenteller Fähigkeiten (Osterhaus et al. 2017)	–	–	–	12	12	2–4	MA + Begründung, MC + Begründung	Alltag
Scientific Thinking Test (Koeber et al. 2015)	–	8	7	4	19	2–4	MC, MA	Alltag, Physik
Zahngesundheit von Croker und Buchanan (2011)	4	–	–	–	4	1–4	Interview	Alltag
CVS Posttest von Dean und Kuhn (2007)	5	–	–	–	5	4	Interaktiver online Test	Physische Geographie
Interventionsstudie von Klahr und Nigam (2004)	1	–	1	–	2	3–4	Interview mit Experimenten	Physik
VKS Prä- und Posttest Posteraufgabe	2	–	–	12	14	3–4	Interview zur Verbesserung von Postern, die Experimente darstellen	Physik, Alltag
Interventionsstudie von Chen und Klahr (1999)	3	–	3	3	9	2–4	Interview mit Experimenten	Physik
VKS Prä- und Posttest VKS Transfertest	–	15	–	5	20	2–4	MC	Biologie Alltag
Flugzeugaufgabe von Bullock und Ziegler (1999)	1	1	–	1	3	3–4	Interview mit Auswahl von Experimenten auf Abbildungen	Alltag
Laternenaufgabe von Bullock (1991)	1	1	1	–	3	2–4	Interview mit Auswahl von Experimenten auf Abbildungen	Alltag
Aufgaben von Tschirgi (1980)	–	8	–	–	–	2 & 4	Interview mit Auswahl von Experimenten auf Abbildungen	Alltag
Piaget Interview (Inhelder und Piaget 1958)	10	–	–	–	10	1–4	Interview mit Experimenten	Physik, Chemie

VKS Variablenkontrollstrategie, *PL* Planung kontrollierter Experimente, *ID* Identifizierung kontrollierter Experimente, *IN* Interpretation der Ergebnisse kontrollierter Experimente, *VER* Verständnis der fehlenden Aussagekraft konfundierter Experimente (Zurückweisung einer kausalen Interpretation konfundierter Experimente), *JG* Jahrgangsstufen in denen der Test eingesetzt wurde, *MA* Mehrfachauswahl, *MC* Multiple-Choice, *OA* offene Aufgaben

bereich hingegen wurde der Einfluss der Fachkontexte der Aufgaben auf die Strukturierung noch nicht untersucht.

Variablenkontrolle als domänenübergreifende Fähigkeit

Neben dem Fachkontext der Aufgaben stellen Aufgabentypen, zu deren Bearbeitung jeweils andere Teilfähigkeiten der VKS benötigt werden, mögliche weitere Kategorien zur Strukturierung der VKS dar. Chen und Klahr (1999) nennen insgesamt vier typische Teilfähigkeiten im Zusammenhang mit der experimentellen Hypothesenprüfung, die sie in prozessbezogene und logische Teilfähigkeiten unterteilen. Die prozessbezogenen Teilfähigkeiten fokussieren auf die Planung kontrollierter Experimente bzw. die Identifizierung kontrollierter Vergleiche aus einer Auswahl an kontrollier-

ten und unkontrollierten Experimenten. Auch wenn es sich hierbei jeweils um Ergebnisse der Planung und Identifizierung innerhalb eines validen Experiments handelt, so unterscheiden sich die Teilfähigkeiten in ihrer Komplexität. Bei der Identifikation kontrollierter Experimente müssen gegebene Ausprägungen der unabhängigen Variable verglichen werden, wodurch keine aktive Konstruktion erforderlich ist und die Anzahl möglicher Experimente im Vergleich zur eigenständigen Planung eines Experiments geringer ist. Zu den logischen Teilfähigkeiten zählen die Interpretation der Ergebnisse kontrollierter Experimente sowie die Zurückweisung der Interpretation von unkontrollierten Experimenten. Auch die logischen Problemstellungen unterscheiden sich in ihrer Komplexität. So muss bei der Interpretation der Ergebnisse von kontrollierten Experimenten nur auf Unter-

schiede in der abhängigen Variablen geachtet werden. Um eine Fehlinterpretation von unkontrollierten Experimenten zu vermeiden, muss hingegen zusätzlich kontrolliert werden, dass sämtliche Variablen außer die untersuchten Variable denselben Wert haben. Trotz der unterschiedlichen Fokussierung und Komplexität der vier Teilfähigkeiten basieren alle auf Teilschritten der VKS.

Die unterschiedlichen Teilfähigkeiten sind weitere mögliche Kategorien für die Strukturierung von Aufgaben zur VKS. Dabei ist sowohl eine Strukturierung in zwei Faktoren (prozessbezogene und logische Teilfähigkeiten) als auch in die vier Teilfähigkeiten – Planung, Identifizierung, Interpretation kontrollierter sowie Zurückweisung der Interpretation konfundierter Experimente – denkbar. Für die Sekundarstufe gibt es Studien, die einen Einfluss der Teilfähigkeiten auf die Strukturierung der VKS zeigen. So konnten Schwichow et al. (2016b) und Schwichow et al. (2020) aufzeigen, dass eine Strukturierung nach den Teilfähigkeiten gut zu den bereits gewonnenen Forschungsdaten passt. Bei einer Analyse der VKS bzgl. der Teilfähigkeiten wurde festgestellt, dass die Fähigkeit zum adäquaten Umgang mit konfundierten Experimenten deutlich schwieriger anzuwenden war als die anderen drei Teilfähigkeiten. Erst Schüler*innen der Sekundarstufe II konnten solche Aufgaben lösen. Vor dem Hintergrund dieser Befunde ist die Schweizer Längsschnittstudie, welche die Entwicklung der Fähigkeiten zur VKS im Verlauf der Grundschulzeit untersuchte (2.–6. Jahrgangsstufe), interessant, denn sie wies auf eine vierdimensionale Struktur der VKS hin (Peteranderl und Edelsbrunner 2020).

Befunde zu domänenspezifischen bzw. domänenübergreifenden Eigenschaften der Variablenkontrollstrategie im Grundschulalter

Um einen Überblick über den bisherigen Forschungsstand zu domänenspezifischen bzw. domänenübergreifenden Eigenschaften der VKS in der Grundschule zu erhalten, haben wir Studien aus dem Grundschulbereich systematisch analysiert. Im Fokus unserer Analyse standen die eingesetzten Erhebungsinstrumente, da deren Gestaltung zeigt, inwiefern Studien überhaupt Aussagen über den Einfluss unterschiedlicher Fachkontexte der Aufgaben und Teilfähigkeiten ermöglichen. Dazu haben wir die einzelnen Aufgaben den vier Teilfähigkeiten der VKS zugeordnet und die Fachkontexte und das Format der Testinstrumente kodiert. Die Ergebnisse unserer Analyse werden in Tab. 1 zusammengefasst.

Auf den ersten Blick erstaunt die große Anzahl an Studien zur VKS im Grundschulbereich. Entsprechende Fähigkeiten wurden entweder diagnostiziert oder gefördert, wobei oft identische Testinstrumente zum Einsatz kamen. Dies spiegelt sich darin wider, dass beispielsweise die La-

ternenaufgabe von Bullock (1991) auch von Koerber et al. (2011) oder die Flugzeugaufgabe von Bullock und Ziegler (1999) auch von Grygier (2008) und Sodian et al. (2002) verwendet wurde.

Bezogen auf die Fachkontexte fällt auf, dass trotz der Variation vor allem solche verwendet werden, die alltagsbezogene oder physikalische Konzepte beinhalten, die in der Regel nicht vom Sachunterricht abgedeckt werden. Das bedeutet, dass wenige Themen und Konzepte aus dem Sachunterricht eine Rolle spielen. Ausnahmen sind die Aufgaben zum Magnetismus von Bohrmann (2017), zum Schwimmen von Booten sowie zum Füttern von Kühen bei Edelsbrunner et al. (2018) und zum Schwimmen und Sinken bei Koerber et al. (2015). Aus testtheoretischer Sicht ist eine Minimierung des Einflusses des Vorwissens, z. B. durch Alltagskontexte (vgl. Tschirgi 1980) plausibel (Moosbrugger und Kelava 2007). Allerdings ist auf der Grundlage derartiger Testinstrumente nicht erklärbar, inwiefern die verwendeten Fachkontexte die Schwierigkeit der Aufgaben zur VKS beeinflussen.

Unter der Perspektive der VKS als domänenübergreifende Fähigkeit erfasste lediglich das Instrument von Bohrmann (2017) sämtliche Teilfähigkeiten. Die anderen Untersuchungen beschränkten sich auf eine Auswahl an Teilfähigkeiten. Zudem wurden in allen Testinstrumenten die Fachkontexte zusammen mit den Teilfähigkeiten zwischen den Aufgaben variiert. Mit der Annahme der VKS als eindimensionales Konstrukt (z. B. Bohrmann 2017; Koerber et al. 2010) ist eine solche Variation von Fachkontexten und Teilfähigkeiten nachvollziehbar. Jedoch kann aufgrund der Konfundierung von Fachkontexten und Teilfähigkeiten in den vorliegenden Instrumenten nicht geklärt werden, inwiefern die beiden Kategorien einen Einfluss auf die Strukturierung der VKS haben.

Zusammengefasst zeigen die Ergebnisse unserer Analyse, dass bisherige Testinstrumente mit nur einer Ausnahme nicht sämtliche Teilfähigkeiten erfassen und dass die Fachkontexte der Aufgaben innerhalb der Instrumente nur wenig variieren bzw. diese nicht relevant für den Sachunterricht sind. Zudem werden innerhalb der Instrumente die Fachkontexte der Aufgaben zusammen mit den Teilfähigkeiten variiert, so dass ihr jeweiliger Einfluss auf die Struktur der VKS nicht getrennt beschrieben werden kann. Um diese Forschungslücke zu schließen, haben wir ein schriftliches Testinstrument entwickelt, in dem die Teilfähigkeiten Identifizierung und Interpretation kontrollierter Experimente als Vertreter von prozessbezogenen und logischen Aspekten der VKS in denselben fünf grundschulspezifischen Fachkontexten umgesetzt wurden.

Forschungsfragen

Ziel unserer Studie ist die psychometrische Erfassung und Modellierung der VKS in der Grundschulzeit und eine empirische Klärung der Frage, ob es sich bei der Variablenkontrollstrategie um eine domänenspezifische oder eine domänenübergreifende Fähigkeit handelt und wie die Dimensionalität des Konstrukts modelliert werden kann.

Wie diskutiert, wurde in bisherigen Studien nicht ausreichend systematisch untersucht, inwiefern es sich bei der VKS um eine domänenspezifische oder domänenübergreifende Fähigkeit handelt, obwohl aus der Theorie beide Zugänge denkbar wären. Es ist somit möglich, dass verschiedene Fachkontexte einen Einfluss auf die Dimensionalität der Modellierung der VKS haben (z. B. Gut 2012). Darüber hinaus wurde die VKS in bisherigen Studien überwiegend als ein eindimensionales Konstrukt zusammengefasst (z. B. Edelsbrunner und Dablander 2019). Ausgehend von Chen und Klahr (1999) gehen wir im Rahmen unserer Untersuchung jedoch von einem mehrdimensionalen Konstrukt aus. Die Unterteilung der VKS in verschiedene Teilfähigkeiten lässt uns vermuten, dass die Teilfähigkeiten verschiedene Dimensionen des Gesamtkonstrukts darstellen, was Untersuchungen in der Sekundarstufe (Schwchow et al. 2016b, 2020) und der Grundschule (Peteranderl und Edelsbrunner 2020) bereits bestätigt haben. Aus testökonomischen Gründen (vgl. unten) haben wir uns zunächst auf zwei Teilfähigkeiten der VKS (Identifizierung und Interpretation als Vertreter der prozessbezogenen und logischen Aspekte der VKS) und fünf Fachkontexte beschränkt.

Des Weiteren ist es plausibel, dass es innerhalb der einzelnen identifizierten Dimension schwierigkeitserzeugende Merkmale von Aufgaben gibt (z. B. die Ansprache bestimmter Schülervorstellungen oder die konkrete Aufgabenstellung; eine Konkretisierung findet anhand der konstruierten Aufgaben im Rahmen der Beschreibung der Testentwicklung statt), die eine Fähigkeitsstufenbildung innerhalb der Dimension zulassen. Beide Aspekte, Dimensionalität und Schwierigkeit, fassen wir im Folgenden unter dem Begriff „Struktur der VKS“ zusammen. Es ergeben sich die folgenden Forschungsfragen:

- Forschungsfrage 1: Wie kann die Struktur der VKS in der Grundschule modelliert werden?
 - a) Welche empirisch unterscheidbaren Dimensionen weist die VKS auf?
 - b) Welche schwierigkeitserzeugenden Merkmale (VKS gesamt und innerhalb der identifizierten Dimension) und daraus abgeleiteten Fähigkeitsstufen können identifiziert werden?

Wie in den oben dargestellten Querschnittsstudien beschrieben, findet zwischen der zweiten und vierten Jahrgangsstufe eine erhebliche Veränderung der Fähigkeiten zur

VKS statt (vgl. Koerber et al. 2011; Bullock 1991). Es soll untersucht werden, inwiefern derartige Veränderungen in der zu Forschungsfrage 1 identifizierten Modellierung der VKS abgebildet werden. Von Interesse ist hierbei insbesondere, wie sich die Veränderungen in den Fähigkeitsstufen manifestieren. Bezogen auf die Darstellung der Veränderung der VKS ergibt sich daher die folgende Forschungsfrage:

- Forschungsfrage 2: Wie werden Veränderungen der VKS im Verlauf der Grundschulzeit in der identifizierten Modellierung zur Struktur der VKS abgebildet?

Die oben dargestellten Studien zeigten, dass die VKS durch geeignete unterrichtliche Maßnahmen bereits im Grundschulalter gefördert werden kann (vgl. Bohrman 2017; Chen und Klahr 1999; Klahr und Nigam 2004). Es soll untersucht werden, inwiefern mögliche Veränderungen durch eine unterrichtliche Förderung im identifizierten Modell der VKS abgebildet werden können. Hierbei wird auch die Frage untersucht, inwieweit sich der entwickelte Test anfällig für Testwiederholungseffekte zeigt.

- Forschungsfrage 3: Wie werden Veränderungen der VKS nach Durchlaufen einer unterrichtlichen Förderung in der identifizierten Modellierung zur Struktur der VKS abgebildet?

Die Ergebnisse aus Forschungsfrage 2 und 3 wurden in Relation zueinander gestellt, um abschätzen zu können, in welchem Rahmen die erwarteten Verbesserungen durch eine unterrichtliche Förderung der VKS im Vergleich zur Entwicklung ohne spezifische unterrichtliche Förderung in unsere Modellierung beschrieben werden können.

Methoden

Zur Beantwortung der Forschungsfragen wurden zwei Teilstudien durchgeführt. Für Forschungsfrage 1 wurde aus den Daten einer Querschnittsstudie (2. bis 4. Jahrgangsstufe) die generelle Struktur der VKS in der Grundschulzeit modelliert. Die Ergebnisse wurden für eine Darstellung der Entwicklung der VKS ohne spezifische unterrichtliche Förderung in der identifizierten Modellierung genutzt (Forschungsfrage 2). Zur Beantwortung von Forschungsfrage 3 wurde eine unterrichtliche Förderung zur VKS mit dem Ziel untersucht, die erwarteten Veränderungen im vorgestellten Testinstrument abbilden zu können. In diesem Rahmen wurde auch eine Untersuchung zu Messwiederholungseffekten durchgeführt.

Tab. 2 Schematischer Überblick über die 20 entwickelten Aufgaben aus fünf Fachkontexten, zwei Teilfähigkeiten der VKS und zwei Aufgabentypen pro Teilfähigkeit (5 × 2 × 2); Fachkontexte und Aufgabentypen werden weiter unten beschrieben

Dimensionen der VKS	Fachkontexte (und zugrundeliegende Konzepte)				
	Fallen von Fallschirmen (Luftwiderstand)	Schwimmen/Sinken von Booten (Auftrieb)	Wiegen von Dosen (Dichte)	Tiere beobachten (Geschwindigkeit)	Wasserstand (Wasserdrängung)
Teilfähigkeit Identifizierung	Item 1 ID1	ID1	ID1	ID1	ID1
	Item 2 ID2	ID2	ID2	ID2	ID2
Teilfähigkeit Interpretation	Item 3 IN	IN	IN	IN	IN
	Item 4 BE	BE	BE	BE	BE

Stichproben

Teilstudie 1 (Forschungsfragen 1 und 2)

Um die Ausprägung der VKS über die Grundschulzeit zu erfassen, wurden in einer Querschnittsstudie insgesamt 481 Schüler*innen der zweiten bis vierten Jahrgangsstufe (im Folgenden als „Baseline“ bezeichnet) im Rahmen des herkömmlichen Unterrichts befragt. Durch die Verwendung von Testheften mit verschiedenen Aufgaben lagen designbedingt viele Lücken vor, weshalb recht enge Kriterien zum Ausschluss von Fällen aufgrund fehlender Werte angewendet wurden. Nach Entfernung der Fälle, die weniger als 50 % der Aufgaben eines Testhefts und weniger als 25 % der Aufgaben zu einem Fachkontext bearbeitet hatten (Entfernung von 23 aus Jahrgangsstufe 2 (12 % der befragten Schüler*innen aus Jahrgangsstufe 2), 26 aus Jahrgangsstufe 3 (20 %), 17 aus Jahrgangsstufe 4 (9 %)), verblieben 415 Schülerinnen (49,4 %) und Schüler (50,6 %) der Jahrgangsstufe 2 (165, 39,8 %), 3 (105, 25,3 %) und 4 (145, 34,9 %).

Teilstudie 2 (Forschungsfrage 3)

Um zu überprüfen, inwiefern die Veränderung der Ausprägung der VKS durch eine unterrichtliche Förderung durch den Test abgebildet werden kann, wurden zwei dritte Klassen, die eine solche Förderung durchlaufen haben und nicht Bestandteil der Teilstudie 1 waren, mit einem identischen Prä-, Post- und Follow-Up Fragebogen befragt. Erhebungszeitpunkt 1 (Prätestung) fand zwei Tage vor der unterrichtlichen Förderung statt, Erhebungszeitpunkt 2 (Posttestung) circa eine Woche nach der unterrichtlichen Förderung und

Erhebungszeitpunkt 3 (Follow-Up-Untersuchung) circa drei Monate später. Über alle drei Testzeitpunkte lagen 44 vollständige Datensätze von Schülerinnen (41,0 %) und Schülern (59,0 %) vor. Es mussten keine Datensätze entfernt werden. Zur Überprüfung von Testwiederholungseffekten wurde der Test von zwei weiteren dritten Klassen mit insgesamt 29 Schülerinnen (62,1 %) und Schülern (37,9 %) im Abstand von zehn Tagen bearbeitet. Auch hier mussten keine Datensätze entfernt werden.






Testentwicklung



Für die Testentwicklung wurden Erkenntnisse über bisher in der Grundschule benutzte Untersuchungsinstrumente verwendet. Das bisher in der Grundschule benutzte Untersuchungsinstrumentarium umfasst Interviews und schriftliche Tests mit offenen oder geschlossenen Antwortformaten, wobei bereits ab der zweiten Jahrgangsstufe schriftliche Instrumente mit beiden Antwortformaten ihren Einsatz finden (siehe Tab. 1). Erhebungen mit Interviews haben den Vorteil, dass sie weniger stark von Lese- und Schreibfähigkeiten abhängen und auch bei jüngeren Schüler*innen (ab dem Kindergartenalter z. B. bei Croker und Buchanan 2011) verwendet werden können. Ihr Einsatz in Studien mit größeren Stichproben ist jedoch mit einem erheblichen Aufwand verbunden. Ökonomischer ist in solchen Fällen der Einsatz schriftlicher Aufgaben, insbesondere mit geschlossenen Antwortformaten (Moosbrugger und Kelava 2007). Zwar bergen Aufgaben mit geschlossenem Antwortformat die Gefahr der Überschätzung von Schüler*innenleistungen in sich, doch zeigten die Ergebnisse von Koerber et al. (2010) und Pollmeier et al. (2011), dass sie ähnliche Er-



Tab. 3 Fachkontexte der Testhefte und unterrichtlichen Förderung, Zahl gibt die Position des Fachkontexts im Testheft an


	Fallen von Fallschirmen	Schwimmen/Sinken von Booten	Wiegen von Dosen Variante 1	Tiere beobachten	Wasserstand	Wiegen von Dosen Variante 2
Testheft A	1	2	3	–	–	–
Testheft B	–	–	1	2	3	–
Testheft C	2	3	–	–	1	–
Testheft D	3	2	–	1	–	–
Unterrichtliche Förderung	1/2/3	2/3/1	–	–	–	3/1/2



a

Du möchtest herausfinden, wie schnell Schnecken sind.
 Du hast große und kleine Weinbergschnecken (),
 die du über Gras () oder Stein () kriechen
 Sind große Weinbergschnecken () schneller als kleine Weinbergschnecken ()?







Ich lasse nur große Weinbergschnecken kriechen
 
 große Weinbergschnecke auf Gras große Weinbergschnecke auf Stein



Ich lasse große und kleine Weinbergschnecken kriechen
 
 große Weinbergschnecke auf Gras kleine Weinbergschnecke auf Gras



Ich lasse nur eine kleine Weinbergschnecke kriechen

 kleine Weinbergschnecke auf Gras



Ich lasse Weinbergschnecken über Gras und Stein kriechen.
 
 große Weinbergschnecke auf Gras große Weinbergschnecke auf Stein

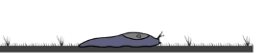
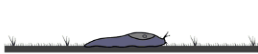
b

Du möchtest erforschen, auf welchem Boden Schnecken schneller kriechen können.
 Du hast Weinbergschnecken () und Nacktschnecken () in groß und klein.
 Die Schnecken können über Gras () oder Stein () kriechen.
 Wie kannst du herausfinden, ob Schnecken auf Gras () oder auf Stein () schneller kriechen? Was musst du vergleichen?



 
 große Weinbergschnecke auf Gras kleine Nacktschnecke auf Stein


 
 große Weinbergschnecke auf Stein kleine Weinbergschnecke auf Gras


 
 große Weinbergschnecke auf Stein große Weinbergschnecke auf Gras

 
 große Nacktschnecke auf Gras große Nacktschnecke auf Gras



c



Du möchtest herausfinden, wie schnell Schnecken unterschiedlicher Art kriechen können.
 Dazu lässt du eine Weinbergschnecke ()
 und eine Nacktschnecke () 10 Zentimeter weit kriechen.




 große Weinbergschnecke
 2 Minuten






 große Nacktschnecke
 5 Minuten

Was zeigt dein Versuch?



Nacktschnecken () sind schneller als Weinbergschnecken ().


Weinbergschnecken () sind schneller als Nacktschnecken ().


Weinbergschnecken () sind genauso schnell wie Nacktschnecken ().

Große Schnecken ( / ) sind schneller als kleine Schnecken ( / ).



d

Du möchtest herausfinden, wie schnell Schnecken unterschiedlicher Art kriechen können.
 Dazu lässt du eine Weinbergschnecke ()
 und eine Nacktschnecke () 10 Zentimeter weit kriechen.




 große Weinbergschnecke
 2 min


 große Nacktschnecke
 5 min

Woher weißt du, was dieser Versuch zeigt?

Ich habe die Kriechzeiten von Weinbergschnecken () und Nacktschnecken () verglichen.

Ich wusste, wie schnell Schnecken kriechen können.

Ich habe die Kriechzeiten von Weinbergschnecken () und Nacktschnecken () verglichen und alles andere gleich gelassen.

Ich habe selber schon Schnecken kriechen lassen.

Abb. 1 Beispielaufgaben zum Fachkontext „Tiere beobachten“ **a** Teilfähigkeit Identifizierung (ID1), **b** Teilfähigkeit Identifizierung (ID2), **c** Teilfähigkeit Interpretation (IN), **d** Teilfähigkeit Interpretation (BE)

gebnisse liefern wie Interviews. Wir haben uns daher entschieden, das Instrument als schriftlichen Test mit geschlossenem (Multiple-Choice) Antwortformat zu konzipieren.

Bei der Testentwicklung wurden zwei der vier Teilfähigkeiten der VKS sowie fünf Fachkontexte berücksichtigt, um die Anzahl der möglichen Aufgaben für eine praktikable Testung (hinsichtlich Testzeit, Anzahl der Testhefte, benötigte Stichprobengröße) in einem angemessenen Rahmen zu halten (Überblick: Tab. 2). Aus den vier oben beschriebenen Teilfähigkeiten entschieden wir uns für Identifizierung und Interpretation, da diese Teilfähigkeiten sowohl prozessbezogene als auch logische Aspekte der VKS berücksichtigen. Außerdem wurden mit diesen Aufgabenformaten in der Grundschule (vorgelagerte Pilotierung siehe unten) und der Sekundarstufe (Brandenburger und Mikelskis-Seifert 2019) gute Erfahrungen hinsichtlich der Praktikabilität gemacht.

Über die zwei Teilfähigkeiten in fünf Fachkontexten haben wir einen schriftlichen Test im Multiple-Choice-Antwortformat entwickelt, der auf dem Test von Schwichow et al. (2016b) basierte. Bei der Testentwicklung wurden ausschließlich kurze Texte und Abbildungen verwendet, um den Einfluss der Lesefähigkeit zu minimieren. In einer vorgelagerten Pilotierung (2.–4. Jahrgangsstufe, elf Klassen, $N=329$) wurde die Verständlichkeit der Aufgaben geprüft, indem Feedback von erfahrenen Grundschullehrkräften eingeholt und eine Strukturierung über die Teilfähigkeiten mit Rasch-Modellierungen erkundet wurde. Inadäquate Aufgaben wurden optimiert bzw. ausgeschlossen.

Der überarbeitete Test beinhaltete die folgenden, grundschulnahen Fachkontexte: Fallen von Fallschirmen, Schwimmen/Sinken von Booten, Wiegen von Dosen, Tiere beobachten und Wasserstand. Die Kontexte der Experimente waren als Fachkontexte so gewählt, dass sie kein spezifisches Vorwissen erforderten, aber einen Bezug zu physikalischen Konzepten aufwiesen. Für jede der untersuchten Teilfähigkeiten wurden zwei Aufgaben verwendet. Um die Anzahl der Aufgaben und die Testzeit in einem verträglichen Maß zu halten sowie die Verzerrung der Ergebnisse, insbesondere durch Reihenfolgeeffekte, abzumildern, wurde ein Testheftdesign verwendet, das eine Variante des „Incomplete Block Designs“ (für eine Übersicht: Frey et al. 2009) darstellt. Hierdurch wurde erreicht, dass sich die Reihenfolge der Fachkontexte änderte und sich die Kontexte zwischen den Testheften überlappten, um die Personenanzahl pro Kontext insbesondere bei den Kontexten zu erhöhen, die auch in der unterrichtlichen Förderung als „Ankeraufgaben“ eingesetzt wurden (siehe Tab. 3). Für die Untersuchung der unterrichtlichen Förderung wurde die Reihenfolge für Prä-, Post- und Follow-Up-Testung variiert. Abb. 1 zeigt Beispielaufgaben aus dem Fachkontext „Tiere beobachten“, wobei es aus physikalischer Sicht um

das Konzept „Geschwindigkeit“ geht. Die Testzeit betrug 45 min.

Die Aufgaben zur Teilfähigkeit Identifizierung (im Folgenden als ID1 und ID2 bezeichnet) konfrontierten die Testpersonen mit einer Frage zur kausalen Wirkung einer unabhängigen Variable (UV) auf eine abhängige Variable (AV). Anschließend mussten die Testpersonen ein passendes kontrolliertes Experiment auswählen. Die beiden Aufgabentypen, die auf Basis der Literatur zu Schülervorstellungen (vgl. Carey et al. 1989; Siler und Klahr 2012; Schwichow et al. 2022) konstruiert wurden, unterschieden sich in der Wahl der Distraktoren: Im Falle der ID1-Aufgaben gab es zwei Antwortoptionen, die aus einem Telexperiment bestanden. Bei den ID2-Aufgaben gab es dagegen einfach konfundierte Experimente (zusätzlich zur UV variierte eine weitere Variable zwischen beiden Bedingungen) und zweifach konfundierte Experimente (zusätzlich zur UV variierten zwei weitere Variablen zwischen beiden Bedingungen). Ergänzt wurden die Antwortoptionen in beiden Aufgabentypen durch nicht-contrastive Experimente und Experimente, in denen eine falsche UV kontrolliert wurde. Zusammengefasst unterschieden sich die Aufgaben zu ID1 und ID2 im Wesentlichen durch die angebotenen Distraktoren. Es wurden zwei Aufgaben konstruiert, um eine größere Bandbreite an Distraktoren und somit an Schülervorstellungen abdecken zu können.

Zur Teilfähigkeit Interpretation wurden für jeden Fachkontext zwei Aufgaben konstruiert (im Folgenden als IN und BE bezeichnet), die auf unterschiedliche Aspekte der Teilfähigkeit eingingen. Die Testpersonen wurden in beiden Aufgaben mit dem identischen grafisch und schriftlich dargestellten Ausgang eines kontrollierten Experiments mit Nennung der AV und UV konfrontiert. Bei den IN-Aufgaben wurde die Interpretation des Experiments abgefragt. Die vier Antwortoptionen deckten einen positiven, negativen bzw. keinen Einfluss der UV und den Verweis auf eine nicht variierte Variable ab. Abhängig vom gezeigten Experiment war entweder ein positiver oder ein negativer Einfluss der UV die richtige Antwort. Die BE-Aufgaben erforderten eine Begründung der Interpretation. Die korrekte Antwort war, dass zwei Telexperimente verglichen wurden, die sich nur in einer UV unterschieden und ansonsten gleiche Ausprägungen der Variablen hatten. In einer weiteren Antwortoption fehlte der Verweis auf die Kontrolle der Variablen. Zudem konnten die Schüler*innen mit eigenem Wissen bzw. eigenen Erfahrungen argumentieren. Zusammengefasst unterschieden sich die Aufgaben IN und BE darin, welcher Teillösungsschritt für eine erfolgreiche Interpretation durchgeführt werden musste.

Die Reihenfolge der Aufgaben innerhalb eines Fachkontextes wurde nicht variiert, sie war immer: ID1, IN, BE, ID2. Bei der Wahl der Reihenfolge wurde berücksichtigt, dass für die Teilfähigkeit Interpretation in den Aufgaben

zu BE die Antworten von IN genannt wurden und so bei einer Umkehr der Reihenfolge der Aufgaben „Carryover Effekte“ auftreten würden (vgl. Frey et al. 2009). Eine Abmilderung dieser Effekte in der Teilfähigkeit Identifizierung sollte dadurch erreicht werden, dass ID1 und ID2 nicht direkt aufeinander folgten.

Unterrichtliche Förderung

Die unterrichtliche Förderung zur VKS basierte auf Fördermaßnahmen, die sich aus der Meta-Analyse von Schwichow et al. (2016a) ergeben haben. Das Durchführen eines Demonstrationsexperiments stellte demnach eine höchst lernwirksame Maßnahme zur Vermittlung der VKS dar. Da weitere Studien zur Lernwirksamkeit von Experimenten (vgl. Meta-Analyse von Ross 1988) jedoch keine signifikanten Unterschiede in der Wirksamkeit im Vergleich von Demonstrations- und Schüler*innenexperimenten zeigten, haben wir bei der Förderung auf beide Experimentierformen zurückgegriffen, da in beiden Fällen ein positiver Effekt auf das Lernen der VKS erwartet wird.

Weiterhin wurde das Auslösen eines kognitiven Konflikts hinsichtlich der VKS als besonders lernwirksame Instruktionsstrategie identifiziert. Dazu wurden Schüler*innen mit konfundierten Experimenten konfrontiert, in denen neben der unabhängigen Variable eine weitere Variable variiert wurde. Entsprechend wirkte sich insbesondere die Kombination aus Demonstrationen und Auslösung eines kognitiven Konfliktes besonders positiv auf den Lernerfolg aus. Diese Kombination aus kognitivem Konflikt und Demonstrationsexperiment wurde bei der zweiphasigen unterrichtlichen Förderung umgesetzt.

Phase 1: Inhalt einer 90-minütigen Lerneinheit waren zwei Schüler*innenexperimente, die in Partner*innenarbeit durchgeführt wurden. Thema des ersten Schüler*innenexperimentes war das Brennen einer Kerze. Ziel dieses Schülerexperimentes war, an das intuitive Verständnis der Schüler*innen von fairen Vergleichen anzuknüpfen und es auf das Konzept kontrollierter Experimente in einen bekannten Fachkontext zu übertragen. Um den Transfer des Konzepts kontrollierter Experimente anzuregen, wurde ein zweites Experiment zu einem für die Schüler*innen unbekanntem Thema (Schwimmen und Sinken) angewendet. Dieses zweite Schüler*innenexperiment diente dazu, den Transfer der VKS anzuregen, indem die Schüler*innen die VKS in einem anderen Fachkontext explizit anwenden mussten.

Phase 2: In der anschließenden 45-minütigen Lerneinheit erfolgte eine explizite Diskussion zur adäquaten Anwendung der VKS, indem bei den Schüler*innen ein kognitiver Konflikt durch ein Demonstrationsexperiment provoziert wurde. Im Anschluss wurde eine weitere Aufgabe zum Anwenden der VKS im Plenum bearbeitet und besprochen.

Datenanalyse

Im Folgenden werden wir, in Anlehnung an die Empfehlungen von Edelsbrunner und Dablander (2019), die Gründe für die Modellwahl, geplante Analyseschritte zur Prüfung der Modellstruktur und Fit-Untersuchungen darstellen.

Zur Modellierung der Struktur der VKS (Forschungsfrage 1) wurden die Daten mit Rasch-Modellen analysiert. Die Rasch-Modelle ermöglichten, die verwendeten Aufgaben und untersuchten Personen hinsichtlich der angesprochenen Fähigkeitsdimension und Aufgabenschwierigkeit zu strukturieren, da beide Kriterien auf derselben Skala abgebildet werden (Boone et al. 2014; Kauertz 2014; Rost 2004). Dies erlaubte zudem die Bildung von Fähigkeitsstufen, was im Hinblick auf die Abbildung von Veränderungen über Jahrgangsstufen und in Folge einer unterrichtlichen Förderung von Interesse ist (Forschungsfrage 2 und 3). Neben inhaltlichen Aspekten sprach auch das Studiendesign für die Verwendung eines Rasch-Modells. Durch die Verwendung unterschiedlicher Testhefte ergaben sich zwangsläufig Lücken im Datensatz, mit denen in einer Rasch-Modellierung gut umgegangen werden konnte, da Überlappungen durch „Ankeraufgaben“ vorlagen (Rost 2004).

Wie zuvor (Forschungsfrage 1) dargestellt, erwarteten wir, dass die VKS als ein mehrdimensionales Konstrukt latenter Fähigkeiten von Schüler*innen modelliert werden kann und dass die Aufgabenschwierigkeit in Verbindung zu schwierigkeitserzeugenden Merkmalen steht. Durch die Konstruktion des Tests konnte plausibel angenommen werden, dass die Aufgaben sowohl voneinander unabhängig als auch unterschiedlich schwierig waren. Neben der Abbildung einer latenten Fähigkeit stellt dies eine wesentliche Voraussetzung für die Auswertung mit Hilfe einer Rasch-Skalierung dar (Kauertz 2014; Rost 2004). Mit dem Vergleich von Rasch-Modellen konnte empirisch überprüft werden, inwieweit eine Modellierung der VKS über mehrere Dimensionen tatsächlich eine bessere Passung zu den Daten aufweist als eine eindimensionale Modellierung (Wu et al. 2007). Die Modellvergleiche wurden hierbei aus statistischer Sicht auf Basis der finalen Deviance der Modellschätzung und Informationskriterien durchgeführt. Nichtsdestotrotz standen die theoretischen Überlegungen zur Mehrdimensionalität der VKS im Vordergrund, da die latente Modellierung, wie in anderen Untersuchungen auch, konfirmatorisch eingesetzt wurde (vgl. Edelsbrunner und Dablander 2019).

Zur Prüfung der Passung der Aufgaben auf die identifizierte Modellstruktur wurden die Aufgaben hinsichtlich ihres Fits (wMNSQ) geprüft und ggf. aus weiteren Analysen ausgeschlossen. Es wurde zudem die Reliabilität der Personenfähigkeit untersucht, wobei zu beachten ist, dass insbesondere bei einer kleinen Anzahl an Aufgaben die Re-

liabilität grundsätzlich niedriger ausfällt (vgl. Lüdtke und Robitzsch 2017).

Die Berechnungen erfolgten mit dem R-Paket TAM (MML 1PL; Robitzsch et al. 2021). Die Aufgabenschwierigkeiten wurden auf 0 normiert, da für spätere Auswertungen (Forschungsfrage 3) die Personenwerte der Schüler*innen aus der unterrichtlichen Förderung nachträglich vor dem Modell der Baseline eingeordnet wurden. Als Schätzer für die Personenfähigkeit wurden für jede Person 20 Plausible Values (PVs) aus der erwarteten a-posteriori-Verteilung (EAP) gezogen. Viele Large Scale Untersuchungen (z.B. Wu 2005) greifen auf fünf PVs zurück. Mehr Ziehungen führen jedoch zu genaueren Schätzungen bei Verfahren zum Umgang mit fehlenden Werten wie PVs (z.B. Bodner 2008; Graham et al. 2007). Zwanzig Ziehungen waren ein Kompromiss zwischen der Vermeidung von zusätzlichem Rechenaufwand und möglichst hoher Genauigkeit der Ergebnisse. Die Verwendung von Plausible Values erlaubte im Vergleich zu EAP oder WLE eine um Messfehler bereinigte Schätzung auf Populationsebene (vgl. Lüdtke und Robitzsch 2017), ist jedoch nicht als Individualschätzer geeignet (Rost 2004). Alle Auswertungen wurden so getätigt, dass für jeden PV eine einzelne Berechnung stattfand. Für das Ergebnis wurde der Mittelwert aus den Ergebnissen der 20 einzelnen Berechnungen mit den PVs bestimmt (OECD 2009). Die Gesamtvarianz wurde unter Berücksichtigung der mittleren Varianz aller 20 Ziehungen und der Imputationsvarianz der 20 mittleren Personenwerte gewichtet berechnet (Details siehe OECD (2009): PISA Data Analysis Manual SPSS, S. 118 ff.).

Die Schüler*innen werden über ihren Personenwert Fähigkeitsstufen zugeordnet (Details in den Ergebnissen). Für eine zusammenfassende Angabe der Häufigkeit wurde für jeden PV eine Häufigkeitsverteilung der Stufen berechnet. Die zusammenfassende Angabe der Häufigkeiten in Prozent entspricht den Mittelwerten über alle PVs (Vorgehen siehe OECD 2009).

Als Vorbereitung für die Rasch-Skalierung wurden die Antworten dichotom kodiert, wobei „1“ für eine korrekte Antwort und „0“ für eine falsche steht. Alle Aufgaben wurden hinsichtlich ihrer Lösungswahrscheinlichkeit untersucht (Darstellung siehe Ergebnisse) und eine Aufgabe zu BE wurde aufgrund einer geringen Lösungswahrscheinlichkeit (0,08) entfernt.

Ergebnisse

Deskriptive Statistiken

Um einen grundsätzlichen Eindruck vom Abschneiden der Schüler*innen beim Test zu erhalten, wurden die erreichten Gesamtpunkte (Summenscore) aus der 0/1-Kodierung

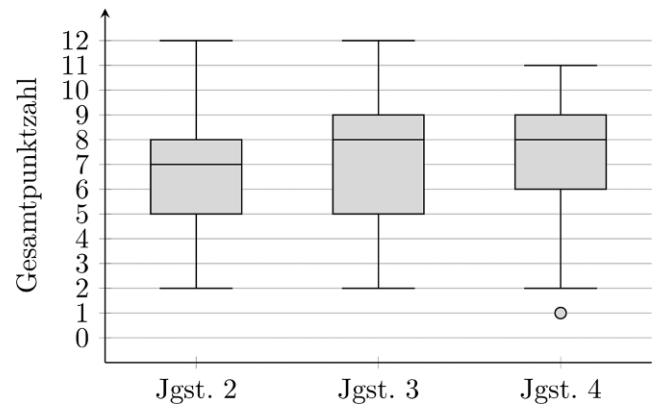


Abb. 2 Boxplot Punkte je Jahrgangsstufe (Jgst.)

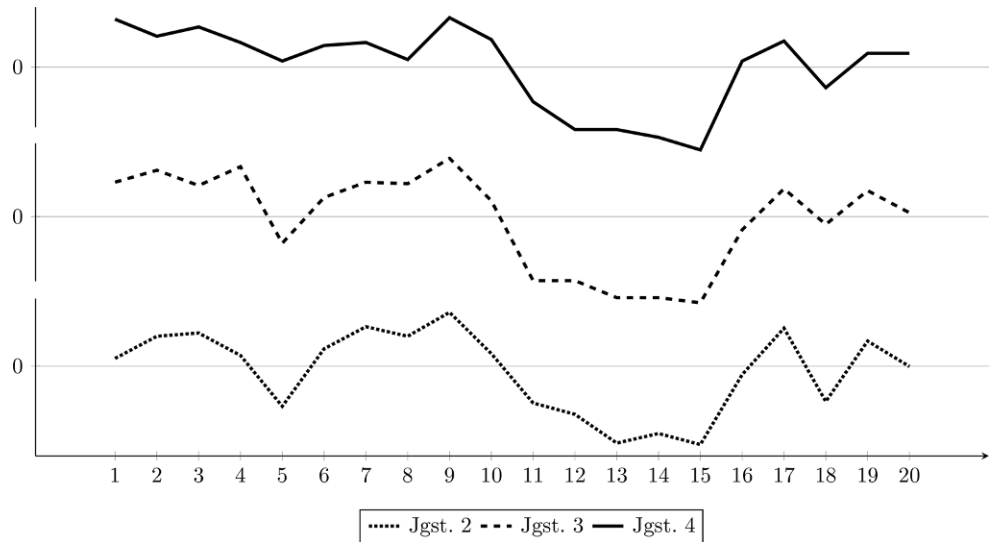
berechnet. Insgesamt wurden im Schnitt 7,01 Punkte von 12 Punkten ($SD=2,25$) erreicht. Abb. 2 zeigt den Boxplot zu den erreichten Punkten in jeder der drei Jahrgangsstufen. Insgesamt streuten die erreichten Punkte in allen Jahrgangsstufen, es zeigten sich keine Decken- oder Bodeneffekte. Eine optische Prüfung der Q-Q-Plots zeigte eine annähernde Normalverteilung mit leichter Rechtsschiefe.

Um zu überprüfen, ob die Aufgaben für die Schüler*innen aller Jahrgangsstufen in Relation zur Gesamtschwierigkeit des Tests vergleichbar gut zu lösen waren, wurden Profile der Lösungswahrscheinlichkeiten erstellt. Damit sollte beispielsweise geprüft werden, ob eine Aufgabe, die für Jahrgangsstufe 2 eher leicht war, auch für Jahrgangsstufe 4 eher leicht war. Hierzu wurde für jede Jahrgangsstufe die Lösungswahrscheinlichkeit der Aufgaben z-transformiert (Bezugspunkt: mittlere Lösungswahrscheinlichkeit der Jahrgangsstufe, d.h. Summenscore durch Aufgabenanzahl) und graphisch aufgetragen (Abb. 3). Es zeigte sich in allen drei Jahrgangsstufen ein ähnlicher Profilverlauf. Das bedeutet, dass die Aufgaben in allen Jahrgangsstufen, relativ zur Gesamtlösungswahrscheinlichkeit, gleich gut gelöst werden konnten und wir so von einer vergleichbaren „Testfairness“ über alle Jahrgangsstufen hinweg ausgehen konnten. So hatte beispielsweise Aufgabe 9 in allen Jahrgangsstufen eine größere Lösungswahrscheinlichkeit als der durchschnittliche Wert für den Test. Aufgabe 13 hingegen hatte eine geringere Lösungswahrscheinlichkeit.

Dimensionsprüfung (Forschungsfrage 1a)

Mit dem Ziel, die Strukturen der VKS zu klären, wurden im ersten Schritt mit den Daten aus der Querschnittsstudie zur Prüfung der *Dimensionalität* verschiedene Rasch-Skalierungen durchgeführt (Forschungsfrage 1a). Neben dem eindimensionalen Grundmodell wurde ein zweidimensionales Modell über die Teilfähigkeiten (Identifizierung, Interpretation), ein vierdimensionales Modell über die Auf-

Abb. 3 Profile der z-transformierten Lösungswahrscheinlichkeiten nach Jahrgangsstufe (Jgst.). Die z-Transformation erfolgte jahrgangweise, damit die Lösungswahrscheinlichkeiten vergleichbar sind. Aus diesem Grund ist der Mittelwert für alle Jahrgangsstufen null. Die Aufgaben sind nach Aufgabentypen geordnet und entsprechen der Reihenfolge im Testheft. 1–5: ID1; 6–10: IN; 11–15: BE; 16–20: ID2



gabentypen (ID1, ID2, IN, BE) und ein fünfdimensionales Modell über die fünf verwendeten Fachkontexte berechnet. Das Modell über die Aufgabentypen diente zur Kontrolle, da die Aufgabentypen z. B. durch die Ansprache verschiedener Schülervorstellungen potenziell auch Dimensionen darstellen könnten. Um die Passung der Modelle zu vergleichen, wurden die Deviance der Modelle unter der Anzahl der Modellparameter einander gegenübergestellt (Tab. 4).

In die Modellauswahl wurden sowohl statistische Kennwerte als auch inhaltliche Überlegungen einbezogen, da statistische Kriterien alleine keine Schlussfolgerungen zu psychologischen Strukturen zulassen, sondern nur in Kombination mit der theoretischen Einbettung interpretiert werden können (vgl. Edelsbrunner und Dablander 2019). Zunächst wurde die Passung der vier berechneten Modelle an die Daten über die finale Deviance ($-2 \cdot \text{Log-Likelihood}$) der Modelle verglichen. χ^2 -Tests zeigten im Vergleich zum Grundmodell eine signifikant bessere Passung des zweidimensionalen ($\chi^2(2) = 42,38, p = 0,000$) und vierdimensionalen Modells ($\chi^2(9) = 84,34, p = 0,000$). Das fünfdimensionale Modell passte nicht signifikant besser als das Grundmodell ($\chi^2(14) = 22,08, p = 0,077$), was gegen eine Dimensionalität bezüglich der Fachkontexte sprach. Des Weiteren wurden Informationskriterien zur Modellauswahl zu Rate gezogen, wobei sich hier ein uneinheitliches Bild zeigte. Der AIC (Akaike Information Criterion;

Deviance $+2 \cdot \text{Parameter}$) des vierdimensionalen Modells war etwas besser als der des zweidimensionalen Modells (Differenz: 62). Der BIC (Bayesian Information Criterion; Deviance $+\ln(N) \cdot \text{Parameter}$) hingegen ordnete, unter der Gewichtung der Personenanzahl, die Passung des zwei- und vierdimensionalen Modells als gleich gut ein (Differenz: 0).

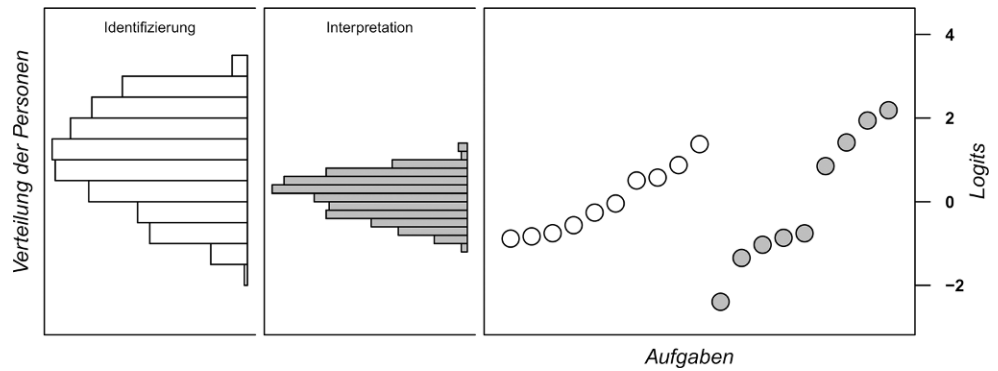
Aus der Perspektive der theoretischen Einbettung war die Wahl des vierdimensionalen Modells, trotz besseren AICs, nicht sinnvoll. Die vier berechneten Dimensionen bildeten sich aus den unterschiedlichen Aufgabentypen: ID1/ID2 unterschieden sich dadurch, dass sie unterschiedliche Distraktoren benutzten, beinhalteten aber ansonsten dieselben Anforderungen an die Teilfähigkeit Identifizierung. IN/BE waren Aspekte der Teilfähigkeit Interpretation, die fest zusammengehörten – eine Interpretation (IN) ist ohne Begründung (BE) nicht sinnvoll. Die Dimensionen des zweidimensionalen Modells stellten die theoretisch verorteten Teilfähigkeiten nach Chen und Klahr (1999) dar. Nach statistischer und inhaltlicher Abwägung entschieden wir uns, nicht zuletzt unter Berücksichtigung des Einfachheitskriteriums nach Rost (2004), für das zweidimensionale Modell (Modellierung der Struktur der VKS über Teilfähigkeiten).

Der Fit (wMNSQ) der Items zum gewählten zweidimensionalen Modell war passend (zwischen 0,8 und 1,2; Neumann 2014), genauso wie die EAP/PV-Reliabilität der Dimensionen (0,65/0,57). Die latente Korrelation der Di-

Tab. 4 Modellvergleiche

N= 415	1-dim. Grundmodell	2-dim. Teilfähigkeiten ID/IN	4-dim. Aufgabentyp ID1/IN/BE/ID2	5-dim. Fachkontext
Deviance	5105,73	5063,35	5021,39	5083,65
Parameter (df)	20	22	29	34
AIC	5146	5107	5079	5152
BIC	5226	5196	5196	5289

Abb. 4 Wright-Map des zwei-dimensionalen Modells nach Teilfähigkeiten (*weiß*: Teilfähigkeit Identifizierung; *grau*: Teilfähigkeit Interpretation)



mensionen war relativ hoch ($r=0,90$), was im Hinblick auf ihre Rollen als Teilfähigkeiten der VKS jedoch plausibel war. Hinsichtlich der Forschungsfrage 1a zeigt sich somit, dass die VKS zwei empirisch unterscheidbare Dimensionen, basierend auf den Teilfähigkeiten Identifizierung und Interpretation, aufwies.

In der Wright-Map (Abb. 4) wurde deutlich, dass die Aufgaben zur Teilfähigkeit Interpretation das Fähigkeitspektrum der Schüler*innen gut abdeckten. Die Aufgaben zur Teilfähigkeit Identifizierung streuten in ihrer Schwierigkeit weniger. Es fehlten im oberen Bereich schwierige Aufgaben.

Deskriptiv betrachtet ergaben sich die in Tab. 5 dargestellten Mittelwerte der Plausible Values (PVs). Hinsichtlich der Personenwerte zeigten sich signifikante Unterschiede zwischen den Jahrgangsstufen mit kleinem Effekt (Identifizierung: $F(2, 412)=5,14, p=0,006, \eta^2=0,02$; Interpretation: $F(2, 412)=3,75, p=0,002, \eta^2=0,01$), was im nächsten Abschnitt über eine Einteilung von Schwierigkeitsstufen genauer untersucht wurde.

Untersuchung der Schwierigkeitsstruktur (Forschungsfrage 1b)

Um zu untersuchen, was die Schwierigkeit der Aufgaben zur VKS beeinflusst (Forschungsfrage 1b), wurden zunächst die Lösungswahrscheinlichkeiten (d.h. Summenscore durch Aufgabenanzahl) verschiedener Aufgabengruppen verglichen. Vergleich man nach *Teilfähigkeit*, zeigten sich Unterschiede bei der mittleren Lösungswahrscheinlichkeit ($t(828)=18,58, p=0,000, d=1,27$). Aufgaben zur Identifizierung ($M=0,75, SD=0,23$) waren im Mittel sig-

nifikant leichter als Aufgaben zur Interpretation ($M=0,45, SD=0,30$). Die unterschiedlichen Schwierigkeiten standen im Einklang mit den verschiedenen Anforderungen der Teilfähigkeiten. Bei der Identifizierung musste nur ein kontrolliertes Experiment ausgewählt werden, wohingegen bei der Interpretation Schlüsse aus kontrollierten Experimenten gezogen und begründet werden mussten.

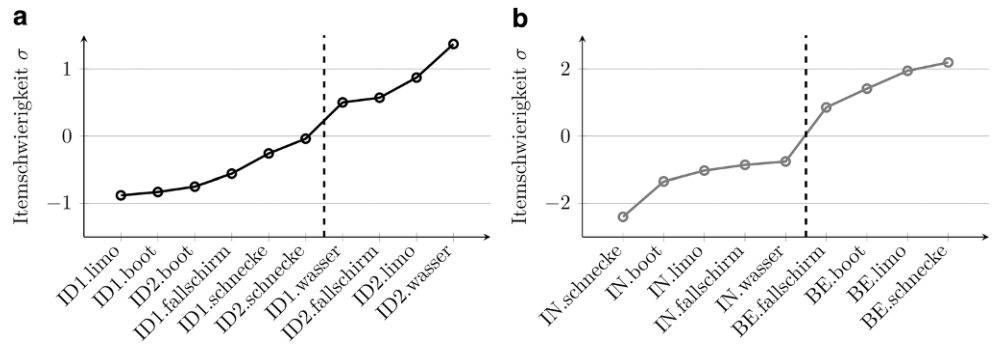
Darüber hinaus wurden die Schwierigkeit innerhalb der identifizierten Dimensionen der Teilfähigkeit untersucht. Es zeigte sich ein deutlicher Einfluss des *Aufgabentyps* auf die Schwierigkeit (Abb. 5).

Teilfähigkeit Identifizierung: Aufgaben, bei denen der Distraktor „nur ein Telexperiment“ verwendet wurde (ID1), waren einfacher als Aufgaben, bei denen der Distraktor „einfach/zweifach konfundiertes Experiment“ verwendet wurde (ID2). **Teilfähigkeit Interpretation:** Aufgaben, bei denen nur interpretiert werden musste (IN), waren einfacher als Aufgaben, bei denen die Interpretation begründet werden musste (BE). Anhand markanter Sprünge in der Schwierigkeit konnten jeweils zwei Fähigkeitsstufen gebildet werden. Es bleibt anzumerken, dass zwei Aufgaben aus der Teilfähigkeit Identifizierung nicht perfekt in die Stufeneinteilung passten (ID2.boot und ID1.wasser). Als Erklärung kann angebracht werden, dass es sich hierbei um eine sehr leichte (ID2.boot) und eine relative schwere Aufgabe (ID1.wasser) handelte. Nichtsdestotrotz galt für beide Aufgaben, dass sie innerhalb eines Fachkontexts zu der grundsätzlichen Tendenz passten und dass ID1-Aufgaben leichter als ID2-Aufgaben waren. Anhand des Personenwerts in den Dimensionen wurden die Schüler*innen für weitere Analysen den identifizierten Stufen zugeteilt. Wie im Methodenteil beschrieben, wurde für jeden PV die Stufe

Tab. 5 Mittlere Personenfähigkeiten in den identifizierten Dimensionen (PV); Berechnung von Mittelwert (M) und Standardabweichung (SD) mit PVs nach OECD (2009)

Dimension	Gesamt		Jahrgangsstufe 2		Jahrgangsstufe 3		Jahrgangsstufe 4	
	M	SD	M	SD	M	SD	M	SD
Identifizierung	1,01	1,31	0,76	1,29	1,17	1,26	1,18	1,32
Interpretation	0,09	0,70	-0,02	0,69	0,17	0,67	0,17	0,71
	$N=415$		$N=165$		$N=105$		$N=145$	

Abb. 5 Schwierigkeitsprofil der Dimensionen. **a** Teilfähigkeit Identifizierung kontrollierter Experimente; Stufengrenze 0,25, **b** Teilfähigkeit Interpretation kontrollierter Experimente; Stufengrenze 0



einer Person bestimmt und über alle PVs der Mittelwert der Häufigkeiten (in Prozent) berechnet.

Im Rahmen der Teilfähigkeit Identifizierung gelang 28,7% der Schüler*innen eine erfolgreiche Identifizierung kontrollierter Experimente nur dann, wenn lediglich Teilexperimente und nicht kontrastive Experimente als Distraktoren angeboten wurden (Stufe 1). 71,3% der Schüler*innen konnten darüber hinaus kontrollierte Experimente auch dann identifizieren, wenn anstelle von Teilexperimenten konfundierte Experimente als Distraktoren verwendet wurden (Stufe 2).

Bei der Teilfähigkeit Interpretation konnten 45,0% der Schüler*innen die Ergebnisse kontrollierter Experimente lediglich interpretieren, jedoch nicht korrekt begründen (Stufe 1). 55,0% der Schüler*innen gelang sowohl eine richtige Interpretation als auch eine korrekte Begründung (Stufe 2).

Hinsichtlich der Forschungsfrage 1b zeigt sich somit, dass zusammenfassend die Teilfähigkeit Identifizierung innerhalb der VKS als einfacher eingeschätzt werden konnte als die Teilfähigkeit Interpretation. Innerhalb der Teilfähigkeit stellte der verwendete Aufgabentyp einen Indikator für die Schwierigkeit dar. Die Schüler*innen konnten hierdurch in jeder der zwei identifizierten Dimensionen einer von zwei Fähigkeitsstufen zugeordnet werden.

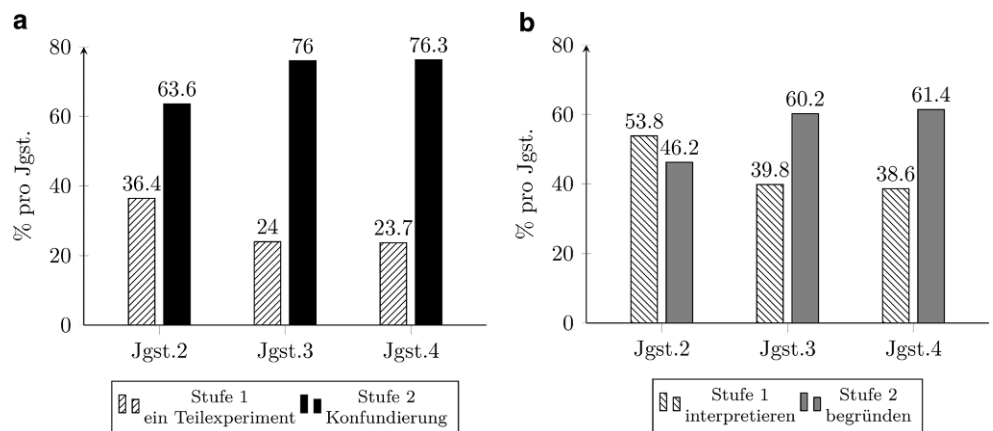
Für die anderen Auswertungen wurde mit der Stufeneinteilung der Schüler*innen weitergearbeitet und nicht mit

den Personenwerten aus der Rasch-Analyse. Wir gingen vor, da die auf den Personenwerten basierenden Stufen inhaltlich über die Stufenbeschreibungen mehr Aussagekraft besitzen als abstrakte Personenwerte.

Darstellung der Veränderung der VKS im Lauf der Grundschulzeit (Forschungsfrage 2)

Im Rahmen der Forschungsfrage 2 wurde untersucht, inwiefern die Veränderungen der VKS im Lauf der Grundschulzeit in der identifizierten Modellierung zur Struktur der VKS abgebildet werden. Beim Vergleich der Verteilung auf die Fähigkeitsstufen zwischen den Jahrgangsstufen (Abb. 6) ließen sich Unterschiede zwischen der Zuteilung auf die Fähigkeitsstufen nach Jahrgangsstufen feststellen (χ^2 -Test Vergleich der Proportionen Teilfähigkeit Identifizierung: $\chi^2(2)=7,66, p=0,022$, Teilfähigkeit Interpretation: $\chi^2(2)=8,71, p=0,013$). Es wurde für beide Teilfähigkeiten ein Trend-Test für die Veränderung von Proportionen durchgeführt (Cochran-Armitage-Trendtest). Sowohl für die Teilfähigkeit Identifizierung ($\chi^2(1)=6,26, p=0,012$) als auch die Teilfähigkeit Interpretation ($\chi^2(1)=7,41, p=0,006$) zeigte sich, dass sich der Anteil der Schüler*innen in den Stufen über die Jahrgangsstufen linear veränderte. Das heißt mit höherer Jahrgangsstufe wurden die Schüler*innen einer höheren Fähigkeitsstufe zugeordnet und konnten demzufol-

Abb. 6 Einteilung Fähigkeitsstufen nach Jahrgangsstufen (Jgst.). **a** Teilfähigkeit Identifizierung, **b** Teilfähigkeit Interpretation. (N=415, Jahrgangsstufe 2: N=165, Jahrgangsstufe 3: N=105, Jahrgangsstufe 4: N=145)



ge zunehmend schwerere Aufgaben erfolgreich bearbeiten. Bei einer genaueren Untersuchung dieses Trends wurde festgestellt, dass sich die Stufenverteilung in Jahrgangsstufe 3 nicht von der in Jahrgangsstufe 4 unterscheiden (Teilfähigkeit Identifizierung: $\chi^2(1)=0,00$, $p=1$, Teilfähigkeit Interpretation: $\chi^2(1)=0,00$, $p=0,951$) und die Veränderung beim Vergleich von Jahrgangsstufe 2 mit Jahrgangsstufe 3 bzw. 4 zu beobachten war. Dieser Befund kann eventuell mit dem Bildungsplan des Landes Baden-Württemberg erklärt werden, der für die Jahrgangsstufen 1 und 2 sowie 3 und 4 gemeinsame Standards festlegt (Ministerium für Kultus, Jugend und Sport Baden-Württemberg 2016).

In Bezug auf Forschungsfrage 2 lässt sich festhalten, dass die Veränderung im Lauf der Grundschulzeit in der identifizierten Modellierung zur Struktur der VKS in den Fähigkeitsstufen abgebildet werden konnte.

Abbildung der Veränderung der VKS im Rahmen einer unterrichtlichen Förderung (Forschungsfrage 3)

Bevor untersucht wurde, inwiefern eine Veränderung der Ausprägung der VKS durch eine unterrichtliche Förderung im identifizierten Modell abgebildet werden kann (Forschungsfrage 3), wurden die deskriptiven Ergebnisse der unterrichtlichen Förderung dargestellt. Um einen Überblick über die Effekte der unterrichtlichen Förderung zu bekommen, wurde zunächst mit Methoden der klassischen Testtheorie verglichen, wie viele Punkte die Schüler*innen zu den Testzeitpunkten erreichten. Der Test bestand aus 12 Aufgaben aus den Fachkontexten „Fallen von Fallschirmen“, „Schwimmen/Sinken von Booten“ und „Dosen wiegen“ (Variante 2), wobei die Ergebnisse einer Aufgabe wegen eines Druckfehlers in der Aufgabenstellung nicht für die Auswertung berücksichtigt werden konnte. Somit waren mit einer 0/1-Kodierung maximal 11 Punkte erreichbar (Abb. 7).

Es konnte ein starker signifikanter Haupteffekt der Testzeitpunkte gemessen werden ($F(2, 131)=16,59$, $p=0,000$, $\eta^2=0,20$). Post-hoc-Tests (Bonferroni) zeigten signifikante

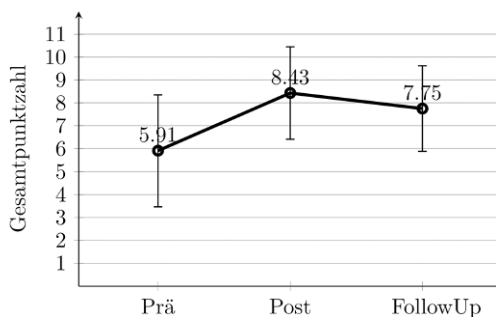


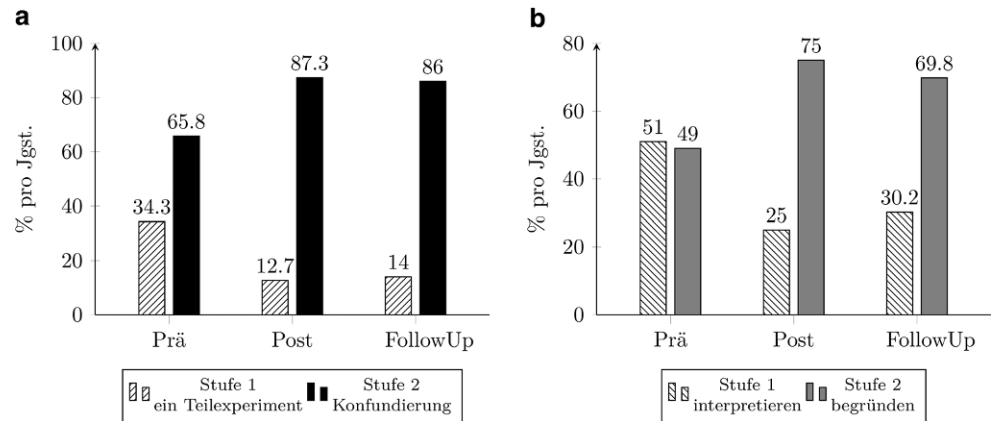
Abb. 7 Vergleich der durchschnittlichen Gesamtpunktzahl zu verschiedenen Testzeitpunkten $N=44$

Unterschiede zwischen Prä- und Post-Test ($p=0,000$) und Prä- und Follow-Up-Test ($p=0,000$). Es zeigten sich keine signifikanten Veränderungen zwischen Post- und Follow-Up-Test ($p=0,405$). Daraus lässt sich schließen, dass die durch eine unterrichtliche Förderung erwartete (stabile) Veränderung der VKS durch das vorgestellte Testinstrument abgebildet werden konnte.

Um die bestätigte Veränderung der VKS durch eine unterrichtliche Förderung in der identifizierten Struktur zur VKS abzubilden (Forschungsfrage 3), wurden die Ergebnisse der unterrichtlichen Förderung in das vorgestellte zweidimensionale Rasch-Modell eingeordnet. Der eingesetzte Test zur Begleitung der unterrichtlichen Förderung wies Übereinstimmungen in sieben Aufgaben zu den Aufgaben der Baseline-Untersuchung auf („Ankeraufgaben“ siehe Tab. 3; eine Aufgabe konnte wegen eines Druckfehlers nicht berücksichtigt werden). Da das Rasch-Modell der Baseline unter der Summenorientierung der Aufgabenschwierigkeiten geschätzt wurde, konnten für die Schüler*innen, die an der unterrichtlichen Förderung teilgenommen haben, auf Basis der sieben übereinstimmenden Aufgaben Personenfähigkeiten berechnet werden, die im gleichen Bezugsrahmen lagen wie die der Baseline (quasi als „fünftes Testheft“ der Baseline mit fehlenden Werten). Hierzu wurde mit den Daten der Baseline und den passenden Daten der unterrichtlichen Förderung eine weitere Rasch-Skalierung durchgeführt, wobei die Aufgabenschwierigkeiten auf die Werte der Skalierung der Baseline festgelegt wurden. Über die Personenfähigkeit (PVs) wurden die Schüler*innen, die an der unterrichtlichen Förderung teilgenommen haben, nach dem gleichen Vorgehen wie oben den identifizierten Fähigkeitsstufen der Dimensionen nach Teilfähigkeiten zugeteilt. Für die Schüler*innen, die an der unterrichtlichen Förderung teilgenommen haben, konnte durch diese Auswertung ein positiver Effekt der unterrichtlichen Förderung im identifizierten Modell zur Struktur der VKS dargestellt und mit der Entwicklung ohne spezifische unterrichtliche Förderung in Relation gesetzt werden.

Es ließen sich Unterschiede zwischen der Zuteilung auf die Fähigkeitsstufe Identifizierung nach Testzeitpunkt feststellen (Vergleich der Proportionen über alle Testzeitpunkte: $\chi^2(2)=7,28$, $p=0,026$). Zum Zeitpunkt der Prä-Testung zeigten die Schüler*innen für die Teilfähigkeit Identifizierung (Abb. 8a) eine etwas schlechtere Verteilung auf die Stufen als eine typische 3. Jahrgangsstufe aus der Baseline (Abb. 6a), das heißt mehr Schüler*innen wurden der Stufe 1 zugeordnet. Nach der unterrichtlichen Förderung wurden nur wenige Schüler*innen noch der Stufe 1 zugeordnet (Vergleich Prä-Post: $\chi^2(1)=4,57$, $p=0,016$, einseitig). 87,3 % der Schüler*innen wurden in der Stufe 2 verortet und konnten nach der unterrichtlichen Förderung bei der Identifizierung von kontrollierten Experimenten auch mit konfundierten Experimenten umgehen. Sie übertrafen

Abb. 8 Einteilung Fähigkeitsstufen im Rahmen der unterrichtlichen Förderung zu den drei Testzeitpunkten, **a** Teilfähigkeit Identifizierung, **b** Teilfähigkeit Interpretation. ($N=44$)



damit die Ergebnisse einer 4. Jahrgangsstufe aus der Baseline (Abb. 6a). Der Effekt wurde auch bei der Follow-Up-Untersuchung stabil dargestellt (Vergleich Post-Follow-Up: $\chi^2(1)=0,00$, $p=1,00$, zweiseitig).

Bei der Teilfähigkeit Interpretation (Abb. 8b) wurden ebenfalls die erwarteten positiven Effekte der unterrichtlichen Förderung durch den Test abgebildet (Vergleich der Proportionen über alle Testzeitpunkte: $\chi^2(2)=7,28$, $p=0,026$). Auch wenn die Schüler*innen zum Zeitpunkt der Prä-Testung in der Verteilung auf die Stufen schlechter abschnitten als eine typische 3. Jahrgangsstufe der Baseline (vgl. Abb. 6b), übertraf die Stufeneinteilung nach der unterrichtlichen Förderung (Vergleich Prä-Post: $\chi^2(1)=5,26$, $p=0,010$, einseitig) die Ergebnisse einer 4. Jahrgangsstufe aus der Baseline (vgl. Abb. 6b). Die Stufenzugehörigkeit blieb auch in der Follow-Up-Testung stabil (Vergleich Post-Follow-Up: $\chi^2(1)=0,09$, $p=0,759$, zweiseitig).

Um abzusichern, dass der eingesetzte Test zur VKS keinen Lerneffekt verursachte (Testwiederholungseffekt), wurde der Prä-Post-Test mit zwei weiteren 3. Klassen ($N=29$) durchgeführt. Hier zeigte sich keine signifikante Veränderung der erreichten Gesamtpunktzahlen ($M_{Prä}=4,21$, $SD_{Prä}=2,37$, $M_{Post}=4,59$, $SD_{Post}=2,21$; $t(56)=0,631$, $p=0,531$). Dieses Ergebnis zeigte sich auch in der Verteilung auf die Stufen. Es ergab sich keine signifikante Veränderung der Stufeneinteilung zwischen Prä- und Post-Test (Vergleich der Proportionen Identifizierung: $\chi^2(1)=0,09$, $p=0,763$; Interpretation: $\chi^2(1)=0,11$, $p=0,737$). Die Ergebnisse legen nahe, dass die Schüler*innen nicht allein durch den Einsatz des Testes zusätzlich Strategien zur Variablenkontrolle lernen.

Diskussion

Die Ergebnisse unserer Studie zeigen, dass sich zwei der aus Chen und Klahr (1999) angenommenen Teilfähigkeiten aus den prozessbezogenen und logischen Aspekten der VKS empirisch als trennbare, wenn auch miteinander hoch

korrelierende Teilfähigkeiten darstellen lassen. Wir haben keinen Einfluss des Fachkontexts der Aufgabe auf die Dimensionalität der VKS identifiziert. Dieser Befund deckt sich mit der getroffenen Annahme, dass die VKS eine domänenübergreifende Fähigkeit ist, die variabel angewendet werden kann. Damit weichen unsere Ergebnisse von den Befunden von Wellnitz et al. (2017) und Gut (2012) ab, die einen Einfluss der Fachkontexte auf die Struktur von Kompetenzen aus dem Bereich der naturwissenschaftlichen Denk- und Arbeitsweisen fanden. Ein augenscheinlicher Unterschied zu den genannten Arbeiten ist, dass unser Instrument ausschließlich auf eine Arbeitsweise (VKS) fokussierte und dass eine Trennung von Teilfähigkeiten und Fachkontexten möglich war. Die genauen Gründe für diese widersprüchlichen Befunde können aufgrund der mannigfaltigen Unterschiede zwischen den Studien nicht hinreichend geklärt werden. Einschränkend muss zudem erwähnt werden, dass wir zwar Fachkontexte mit Bezug zum Sachunterricht gewählt haben, diese jedoch nur auf Konzepten basierten und Variablen enthielten, die aus dem Alltag bekannt sein sollten. Unklar ist, inwiefern es Unterschiede geben könnte, wenn Fachkontexte basierend auf Konzepten, die nicht aus dem Alltag bekannt sind (z.B. Stromstärke, Widerstand), genutzt würden. Um einen tieferen Einblick in die Abhängigkeit der VKS vom jeweiligen Fachkontext zu erhalten, wären somit Studien notwendig, welche systematisch die Fachgebundenheit der gewählten Inhalte variieren, ohne eine Konfundierung mit den Teilfähigkeiten der VKS herzustellen. Grundsätzlich würden Befunde, die auf eine Kontextabhängigkeit hindeuten, jedoch nicht im Widerspruch zur Einschätzung stehen, dass die VKS eine domänenübergreifende Fähigkeit darstellt. Die Einschätzung wird dadurch unterstützt, dass in neueren Konzeptualisierungen des Begriffs „domänenübergreifende Fähigkeit“ davon ausgegangen wird, dass entsprechende Fähigkeiten nicht zwangsläufig ohne das dazugehörige Fachwissen angewendet werden können. Vielmehr ist entscheidend, dass die Fähigkeiten anwendbar sind, und dass Personen, welche über das entsprechende Fachwissen verfügen, dies auch

erkennen und die Fähigkeiten anwenden können (Wellnitz et al. 2017; Hetmanek et al. 2018; Zimmerman 2007).

Keinen Einfluss auf die Dimensionalität hat ferner der Aufgabentyp. Allerdings hatten wir innerhalb der beiden identifizierten Dimensionen (Identifikation und Interpretation) eine klare Abhängigkeit der Aufgabenschwierigkeiten von den Aufgabentypen festgestellt. Bei den Identifizierungsaufgaben waren die Aufgaben, die einen Distraktor enthalten, der nur ein Telexperiment darstellte, einfacher als Aufgaben, die einen weiteren Distraktor mit einem konfundierten Experiment zeigten. Bei den Interpretationsaufgaben waren Aufgaben, die eine reine Interpretation erforderten, einfacher als solche, die zusätzlich eine Begründung einforderten. Dieser Befund steht in keinem Widerspruch zur Dimensionsanalyse. Er bedeutet lediglich, dass die Varianz zwischen Aufgaben unterschiedlicher Dimensionen größer war als die innerhalb der Dimensionen (siehe Wrighth-Map, Abb. 4).

Unsere hier vorgestellte Studie ist nicht die erste Arbeit, welche einen Einfluss der Teilfähigkeiten der VKS nach Chen und Klahr (1999) auf die Struktur der VKS aufzeigt. So berichten bereits Schwichow et al. (2016b, 2020) ähnliche Befunde in zwei Studien mit Schüler*innen in der Sekundarstufe des deutschen Schulsystems. Außerhalb von Deutschland fanden Peteranderl und Edelsbrunner (2020) (Schweiz, Primarstufe) und van Vo und Csapó (2021) (Vietnam, Sekundarstufe) ebenfalls einen Einfluss der Teilfähigkeiten und nicht des Fachkontexts der Aufgabe auf die Struktur der VKS bei Schüler*innen. Aber auch Studien, die sich bei der Aufgabenkonstruktion nicht auf die Definition der Teilfähigkeiten der VKS beziehen, wie beispielsweise Nehring (2015), setzten Aufgaben ein, welche sich inhaltlich den Teilfähigkeiten der VKS zuordnen lassen. Entsprechende Arbeiten beziehen sich auf die traditionelle Einteilung experimenteller Arbeitsweisen in die drei Schritte 1) Hypothesenfindung, 2) Experimentplanung und 3) Auswertung (für eine Übersicht siehe Emden 2011), wobei Aufgaben aus den Bereichen Experimentplanung und Auswertung in ihrer Operationalisierung große Ähnlichkeit in den Teilfähigkeiten der VKS aufweisen. So werden die Testpersonen bei Nehring (2015) bei Aufgaben zum Bereich „Experimentplanung“ aufgefordert, ein geeignetes Experiment zur Testung einer gegebenen Hypothese auszuwählen. Wie bei Aufgaben zur Teilfähigkeit Identifizierung (ID) werden den Schüler*innen als Antwortmöglichkeiten kontrollierte und konfundierte Experimente angeboten. Aufgaben aus dem Bereich „Auswertung“ fordern von den Testpersonen hingegen das gegebene Ergebnis eines Experiments zu interpretieren und dazu einen passenden Antwortsatz auszuwählen. Hier besteht eine Ähnlichkeit zu den Aufgaben der Teilfähigkeit Interpretation (IN).

Implikationen

Über eine Strukturierung der VKS und damit verbunden aus den unterschiedlichen Fähigkeitsstufen innerhalb der Dimensionen lassen sich direkte Konsequenzen für die Messung dieser Strategie ableiten. So scheinen die meisten Schüler*innen zu wissen, dass zum Prüfen einer Hypothese zwei Bedingungen verglichen werden müssen. Aufgaben, die den Distraktor „nur eine Bedingung“ haben, sind daher besonders einfach und eventuell nur für untere Jahrgangsstufen geeignet, da sonst die Gefahr eines Deckeneffekts besteht. Auf der anderen Seite sind Aufgaben, die Begründungen für die Interpretation von Experimenten einfordern, deutlich schwieriger. Aus diesem Grund sind sie vor allem für höhere Jahrgangsstufen und zum Nachweis von Interventionseffekten geeignet, weil schwierigere Aufgaben größere Interventionseffekte aufdecken können (Schwichow et al. 2016a). In einem solchen Zusammenhang wäre es sinnvoll, nicht nur Aufgaben zu unterschiedlichen Dimensionen einzusetzen, sondern die Ergebnisse auch getrennt zu berichten, um ein genaueres Bild von Interventionseffekten zu erhalten.

Wie bereits Bullock (1991) und Koerber et al. (2011) konnten wir in unserem Strukturmodell zur VKS große Veränderungen zwischen der zweiten und vierten Jahrgangsstufe aufzeigen. So stieg der Anteil an Testpersonen, die der zweiten Fähigkeitsstufe zugeordnet wurden, von 64 % auf 76 % beim Identifizieren und von 46 % auf 61 % beim Interpretieren. Die Zahlen für die Teilfähigkeit Interpretieren haben zum Ende der 4. Jahrgangsstufe dieselbe Größenordnung wie in den genannten Studien. Dies ist ein Argument für die Validität unserer Befunde und insbesondere für die Möglichkeit, auch in der Grundschule mit schriftlichen Tests zu ähnlichen Befunden wie mit Interviews zu kommen. Für den Bereich Identifizieren liegen unsere Ergebnisse deutlich über den Ergebnissen der oben genannten Studien. Dies spricht ebenfalls dafür bei der Erfassung der VKS unterschiedliche Teilfähigkeiten zu berücksichtigen, um Veränderungen genauer erfassen zu können.

Der von uns entwickelte Test ist geeignet, um Unterrichtseffekte nachzuweisen. Trotz einer zeitlich beschränkten Förderung der VKS konnte festgestellt werden, dass Drittklässler das Niveau von Viertklässlern erlangen konnten. Dies legt nahe, die VKS gezielt im Unterricht zu fördern. Zwar schien es auch ohne explizite Förderung im Lauf der Schulzeit einen Zuwachs an entsprechenden Fähigkeiten zu geben, doch unsere Ergebnisse offenbarten, dass dieser mit geringem Aufwand vorweggenommen werden kann. Dieser Befund ist insbesondere insofern interessant, da aufgrund der hohen Korrelation zwischen Fachwissensentwicklung und Fähigkeiten zur VKS (Schwichow und Nehring 2018, 2020; Edelsbrunner et al. 2018) der Erwerb von Inhaltswissen positiv beeinflusst werden kann.

Für die Gestaltung von Fördermaßnahmen bezüglich der VKS liefern unsere Befunde ebenfalls Anhaltspunkte. Eine explizite Thematisierung von Experimenten, die nur aus einer Bedingung bestehen, scheint vor allem in der 3. und 4. Jahrgangsstufe nicht notwendig zu sein, weil die meisten Schüler*innen diese Option nicht wählen. Dagegen sollte die Begründung bei der Interpretation von Ergebnissen stärker thematisiert werden, weil diese auch vielen älteren Schüler*innen Schwierigkeiten bereitet.

Limitationen

Auch wenn unsere Befunde zeigen, dass mit einem schriftlichen Test detaillierte Ergebnisse zu Veränderungen der VKS während der Grundschulzeit abgebildet werden können, haben sie Limitationen. So erfassten wir nur zwei der vier Teilfähigkeiten der VKS, um den Testumfang in einem für die Grundschule realistischen Umfang zu halten. Ferner wurden im bisherigen Testinstrument nicht in sämtlichen Aufgaben alle bekannten Schülervorstellungen zur VKS als Distraktoren angeboten. Für ein solches Vorgehen hatten wir uns entschieden, um die gewählten Schülervorstellungen mindestens einmal anzubieten. Kehrseite dieses Vorgehens ist, dass wir mit der bisherigen Version des Tests keine Analyse auf der Ebene von Schülervorstellungen vornehmen konnten, weil eine Konfundierung zwischen Teilfähigkeiten und Präkonzepten vorlag. Bei einer potenziellen Weiterentwicklung des Tests sollten dieser Aspekt berücksichtigt werden, um eine noch präzisere Erfassung von Entwicklungen der VKS zu ermöglichen. Da es sich bei unserer Studie um eine Querschnittsstudie handelt, können wir streng genommen keine Rückschlüsse auf individuelle Entwicklungen ziehen (Forschungsfrage 2), dazu wäre eine Längsschnittstudie nötig gewesen. Nichtsdestotrotz fügen sich unsere Ergebnisse, wie oben dargestellt, sowohl qualitativ als auch quantitativ in durchgeführte Längsschnittstudien (z. B. Bullock 1991) ein. Wie ebenfalls bereits weiter oben diskutiert, ist die aktuelle Testversion auf Variablen beschränkt, die zwar einen Bezug zu Themen und Konzepten des Sachunterrichts haben, die jedoch aus dem Alltag bekannt sein sollten. Somit wird die Frage des Einflusses des Fachkontexts der Aufgaben nur eingeschränkt untersucht. Eine offene Frage ist daher, inwiefern durch Rückgriff auf fachgebundenerer Variablen andere Ergebnisse eintreten.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Bodner, T.E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4), 651–675. <https://doi.org/10.1080/10705510802339072>.
- Bohrmann, M. (2017). *Zur Förderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht*. Dissertation. Berlin: Logos. <https://doi.org/10.5281/zenodo.1069403>.
- Boone, W.J., Staver, J.R., & Yale, M.S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-007-6857-4>.
- Brandenburger, M., & Mikelskis-Seifert, S. (2019). Facetten experimenteller Kompetenz in den Naturwissenschaften. In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe: Tagungsband zur Jahrestagung 2018 der GDGP in Kiel*. 77–80. Regensburg: Gesellschaft für Didaktik der Chemie und Physik.
- Bullock, M. (1991). Scientific reasoning in elementary school: developmental and individual differences. In *SRCD*. Seattle, WA. Paper.
- Bullock, M., & Ziegler, A. (1999). Scientific reasoning: developmental and individual differences. In F.E. Weinert & W. Schneider (Hrsg.), *Individual development from 3 to 12: findings from the Munich longitudinal study* (S. 38–54). Cambridge: Cambridge University Press.
- Bybee, R.W. (1997). *Achieving scientific literacy: from purposes to practices*. Portsmouth, NH: Heinemann.
- Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). „An experiment is when you try it and see if it works“: a study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education*. <https://doi.org/10.1080/0950069890110504>.
- Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098–1120. <https://doi.org/10.1111/1467-8624.00081>.
- Crocker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: the effect of prior belief and outcome on children's hypothesis-testing strategies. *British Journal of Developmental Psychology*, 29, 409–424. <https://doi.org/10.1348/026151010X496906>.
- Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: the long view. *Science Education*, 91(3), 384–397. <https://doi.org/10.1002/sce.20194>.
- Edelsbrunner, P.A., & Dablander, F. (2019). The psychometric modeling of scientific reasoning: a review and recommendations for future avenues. *Educational Psychology Review*, 31(1), 1–34. <https://doi.org/10.1007/s10648-018-9455-5>.
- Edelsbrunner, P.A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: a large-scale quantitative study in elementary school. *Learning and Individual Differences*, 66, 38–53. <https://doi.org/10.1016/j.lindif.2018.02.003>.
- Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens: Eine vergleichende Studie*.

- die zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I. Berlin: Logos.
- Frey, A., Hartig, J., & Rupp, A.A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice*, 28(3), 39–53. <https://doi.org/10.1111/j.1745-3992.2009.00154.x>.
- GDSU (Hrsg.). (2013). *Perspektivrahmen Sachunterricht*. Bad Heilbrunn: Julius Klinkhardt.
- Graham, J.W., Olchowski, A.E., & Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science: the Official Journal of the Society for Prevention Research*, 8(3), 206–213. <https://doi.org/10.1007/s1121-007-0070-9>.
- Grygier, P. (2008). *Wissenschaftsverständnis von Grundschulern im Sachunterricht*. Bad Heilbrunn: Klinkhardt.
- Gut, C. (2012). *Modellierung und Messung experimenteller Kompetenz. Analyse eines large-scale Experimentiertests*. Berlin: Logos.
- Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge: scientific reasoning and argumentation as a set of cross-domain skills. In F. Fischer, C.A. Chinn, K.F. Engelmann & J. Osborne (Hrsg.), *Scientific reasoning and argumentation: the roles of domain-specific and domain-general knowledge* (S. 203–226). New York, London: Routledge.
- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: an essay on the construction of formal operational structures*. London: Routledge and Kegan Paul. <https://doi.org/10.1037/10034-000>.
- Kauertz, A. (2014). Entwicklung eines Rasch-skalierten Leistungstests. In D. Krüger, H. Schecker & I. Parchmann (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 341–353). Berlin: Springer Spektrum. <https://doi.org/10.1007/978-3-642-37827-0>.
- Kirchner, S. (2013). *Der Umgang mit Variablen bei offenen Experimentieraufgaben im Physikunterricht. Eine Beobachtungsstudie am Beispiel der Konstruktion von auftriebserzeugenden Profilen für ein Windradmodell*. Dissertation. Berlin: Humboldt-Universität.
- Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science*, 15(10), 661–667. <https://doi.org/10.1111/j.0956-7976.2004.00737.x>.
- KMK (2015). *Empfehlungen zur Arbeit in der Grundschule: Beschluss der Kultusministerkonferenz*.
- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: a comprehensive inventory. *Child Development*, 86(1), 327–336. <https://doi.org/10.1111/cdev.12298>.
- Koerber, S., Sodian, B., Kropf, N., Mayer, D., & Schwippert, K. (2011). Die Entwicklung des wissenschaftlichen Denkens im Grundschulalter. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, 43(1), 16–21. <https://doi.org/10.1026/0049-8637/a000027>.
- Koerber, S., Sodian, B., Mayer, D., Kropf, N., Schwippert, K., & Möller, A. (2010). Development of scientific reasoning and science understanding in elementary school. In *21st Biennial international congress of the International Society for the Study of Behavioural Development*. Lusaka, Zambia. Paper.
- Koslowski, B. (1996). *Theory and evidence. The development of scientific reasoning* (1. Aufl.). Learning, development, and conceptual change. Cambridge: MIT Press.
- Lüdtke, O., & Robitzsch, A. (2017). Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. *Diagnostica*, 63(3), 193–205. <https://doi.org/10.1026/0012-1924/a000175>.
- Ministerium für Kultus, Jugend und Sport Baden-Württemberg (2016). Bildungsplan der Grundschule. Sachunterricht. http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GS_SU.pdf. Zugegriffen: 8. Febr. 2021.
- Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-20072-4_2.
- Nehring, A. (2015). *Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie: Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung*. Studien zum Physik- und Chemielernen, Bd. 177. Berlin: Logos. Zugl.: Berlin, Humboldt-Univ., Diss., 2014.
- Nehring, A., Stiller, J., Nowak, K.H., Upmeier zu Belzen, A., & Tiemann, R. (2016). Naturwissenschaftliche Denk- und Arbeitsweisen im Chemieunterricht – eine modellbasierte Videostudie zu Lerngelegenheiten für den Kompetenzbereich der Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften*, 22(1), 77–96. <https://doi.org/10.1007/s40573-016-0043-2>.
- Neumann, K. (2014). Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, H. Schecker & I. Parchmann (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 355–369). Berlin: Springer Spektrum. <https://doi.org/10.1007/978-3-642-37827-0>.
- OECD (2009). *PISA data analysis manual SPSS: SPSS* (2. Aufl.). Paris: Organization for Economic Cooperation & Development.
- Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: children’s social cognition and their epistemological understanding promote experimentation skills. *Developmental psychology*, 53(3), 450–462. <https://doi.org/10.1037/dev0000260>.
- Peteranderl, S., & Edelsbrunner, P.A. (2020). The predictive value of children’s understanding of indeterminacy and confounding for later mastery of the control-of-variables strategy. *Frontiers in Psychology*, 11, 531565. <https://doi.org/10.3389/fpsyg.2020.531565>.
- Pollmeier, J., Hardy, I., Koerber, S., & Möller, K. (2011). Lassen sich naturwissenschaftliche Lernstände im Grundschulalter mit schriftlichen Aufgaben valide erfassen? *Zeitschrift für Pädagogik*, 57(6), 834–853. <https://doi.org/10.3262/ZP1106834>.
- Robitzsch A., Kiefer T., Wu M. (2021). *TAM: Test Analysis Modules*. R package version 3.7-16, <https://CRAN.R-project.org/package=TAM>. Zugegriffen: 22. März 2022.
- Ross, J.A. (1988). Controlling variables: a meta-analysis of training studies. *Review of Educational Research*, 4, 405–437. <https://doi.org/10.3102/00346543058004405>.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern, Göttingen: Huber.
- Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016a). Teaching the Control-of-Variables Strategy: A Meta Analysis. *Developmental Review*, 39, 37–63. <https://doi.org/10.1016/j.dr.2015.12.001>.
- Schwichow, M., Christoph, S., Boone, W.J., & Härtig, H. (2016b). The impact of sub-skills and item content on students’ skills with regard to the control-of-variables-strategy. *International Journal of Science Education*, 38(2), 216–237. <https://doi.org/10.1080/09500693.2015.1137651>.
- Schwichow, M. & Nehring, A. (2018). Variablenkontrolle beim Experimentieren in Biologie, Chemie und Physik: Höhere Kompetenzausprägungen bei der Anwendung der Variablenkontrollstrategie durch höheres Fachwissen? Empirische Belege aus zwei Studien. In: *Zeitschrift für Didaktik der Naturwissenschaften*, 24(1), 217–233. <https://doi.org/10.1007/s40573-018-0085-8>.
- Schwichow, M., Osterhaus, C., & Edelsbrunner, P.A. (2020). The relation between the control-of-variables strategy and content knowledge in physics in secondary school. *Contemporary Educational Psychology*, 63, 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>.
- Schwichow, M., Brandenburger, M., & Wilbers, J. (2022). Analysis of experimental design errors in elementary school: How do students

- identify, interpret, and justify controlled and confounded experiments? *International Journal of Science Education*. <https://doi.org/10.1080/09500693.2021.201554>.
- Siler, S. A., & Klahr, D. (2012). Detecting, classifying and remediating: children's explicit and implicit misconceptions about experimental design. In R. W. Proctor & E. J. Capaldi (Hrsg.), *Psychology of science: implicit and explicit processes* (S. 137–180). New York: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199753628.003.0007>.
- Sodian, B., Thoermer, C., Kircher, E., Grygier, P., & Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule. *Zeitschrift für Pädagogik*, 45. Beiheft, 192–206. <https://doi.org/10.25656/01:3947>.
- Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Development*, 51(11), 1–10. <https://doi.org/10.2307/1129583>.
- Viefers, R., Theyßen, H., & Schreiber, N. (2018). Experimentelle Fähigkeiten in der Grundschule diagnostizieren und individuell fördern. In *PhyDid B Beiträge zur DPG Frühjahrstagung* (S. 277–284).
- van Vo, D., & Csapó, B. (2021). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: evidence of validity and latent predictors of item difficulty. *International Journal of Science Education*, 14(2), 1–21. <https://doi.org/10.1080/09500693.2021.1957515>.
- Wellnitz, N., Hecht, M., Heitmann, P., Kauertz, A., Mayer, J., Sumfleth, E., & Walpuski, M. (2017). Modellierung des Kompetenzbereichs naturwissenschaftliche Untersuchungen. *Zeitschrift für Erziehungswissenschaft*, 20(4), 556–584. <https://doi.org/10.1007/s11618-016-0721-3>.
- Woodward, J. (2003). *Making things happen: a theory of causal explanation*. New York: Oxford University Press. <https://doi.org/10.1093/0195155270.001.0001>.
- Wu, M. L. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2/3), 114–128. <https://doi.org/10.1016/j.stueduc.2005.05.005>.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: generalised item response modelling software*. Camberwell: ACER.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27(2), 172–223. <https://doi.org/10.1016/j.dr.2006.12.001>.