Unterrichtswiss (2021) 49:57–90
https://doi.org/10.1007/s42010-021-00097-9

THEMENTEIL

Evidenzbasiertes Argumentieren bei multivariablen Kausalzusammenhängen mit Interaktionen in der fünften und sechsten Klassenstufe

Sonja Peteranderl · Peter Adriaan Edelsbrunner · Anne Deiglmayr

Eingegangen: 14. August 2020 / Überarbeitet: 21. Januar 2021 / Angenommen: 21. Januar 2021 / Online publiziert: 10. Februar 2021

Zusammenfassung In dieser Studie wird untersucht, in welchem Ausmaß Schüler*innen der 5. und 6. Klassenstufe in einem „Interaktionseffekte"-Kontext, in dem zwei unabhängige Variablen systematisch miteinander interagieren, in ihrer Argumentation vorgegebene experimentelle Evidenz nutzen. Zusätzlich wird untersucht, ob ein im „Haupteffekte"-Kontext situiertes Training der Variablenkontrollstrategie (VKS) die evidenzbasierte Argumentation in einem „Interaktionseffekte"-Kontext beeinflusst. Etwa die Hälfte einer Stichprobe von $N = 618$ Schüler*innen ($M_{\text{alter}} = 11.61$, SD $= 0.65$; 50 % weiblich) erhielt ein im „Haupteffekte"-Kontext situiertes Training der VKS und die andere Hälfte ein aktives Kontrolltraining. Vor und nach dem Training bearbeiteten die Schüler*innen zwei Aufgaben in einem „Interaktionseffekte"-Kontext, in welchem sie ihre Interpretationen präsentierter Evidenz in offenen Antworten begründeten. Zur Klassifizierung des Grades an Evidenzbasierung in den Antworten der Schüler*innen wurde ein Kodiersystem mit fünf Kategorien entwickelt. Analysen der kodierten Antworten zeigen, dass im Vortest bereits etwa 20 Prozent der Schüler*innen in ihren Begründungen den Interaktionseffekt korrekt interpretierten. Der Grad an Evidenzbasierung war bei Schüler*innen der 6. Klassenstufe höher ausgeprägt als bei Schüler*innen der 5. Klassenstufe. Bei einer vorwissensnahen Aufgabe zeigten die Schüler*innen überproportional viele Antworten mit vorwissenbasierter Argumentation ohne Evidenzbezug. Eine kategorienspezifische, ordinale Mehrebenen-Regressionsanalyse zeigte, dass das Training der VKS die Häufigkeit von Begründungen erhöhte, in denen die Schüler*innen Evidenz korrekt nutzten, um Haupteffekte zu interpretieren; auf die Häufigkeit, mit der Interaktionseffekte erkannt wurden, wurde keine Auswirkung gefunden.

S. Peteranderl (✉) · P. A. Edelsbrunner
Research on Learning and Instruction, ETH Zürich, Clausiusstr. 59, 8092 Zürich, Schweiz
E-Mail: sonja.peteranderl@ifv.gess.ethz.ch

A. Deiglmayr
Universität Leipzig, Leipzig, Deutschland

Schlüsselwörter Variablenkontrollstrategie (VKS) · Training · Wissenschaftliches Denken · Evidenzbasierte Argumentation

Evidence-based argumentation in contexts with multivariable causality and interactions in fifth and sixth grade students

Abstract In this study, we examined to which extent within an "interaction effects"-context, in which two independent variables interact systematically, 5th- and 6th-graders consider presented experimental evidence in their argumentations. In addition, we examined whether a training of the control-of-variables strategy (CVS) that is situated within a "main effects"-context influences students' evidence-based argumentation within an "interaction effects"-context. About half of $N = 618$ students ($M_{age} = 11.61$, SD = 0.65; 50% female) from grades 5 and 6 received a training of the CVS situated within the "main effects"-context, and the other half an active control training. Before and after the training, the students worked on two tasks with an "interaction effects"-context in which they justified their interpretations of evidence in open answers. We developed a coding scheme with five categories to classify the level of evidence-based argumentation in students' answers. Analyses of the coded answers showed that, at pretest, already about 20% of students correctly interpreted the interaction effect. The level of evidence-based argumentation was higher among 6th graders than among 5th graders. On a task that was more closely related to students' prior knowledge, students showed an increased number of argumentations that were not based on the presented evidence, but on prior knowledge. A category-specific multilevel ordinal regression analysis showed that the CVS training let to an increase in the frequency of argumentations in which students correctly used evidence to interpret main effects but not to an increase in the frequency with which interaction effects were identified.

Keywords Control of variables strategy (CVS) · Training · Scientific thinking · Evidence-based argumentation

Wissenschaftliches Denken beinhaltet die Fähigkeit zur experimentellen Hypothesentestung: Hierzu gehört, überprüfbare Hypothesen zu formulieren, experimentell zu testen und die so gewonnenen Daten in Bezug auf die zu testende Hypothese korrekt zu interpretieren (Klahr 2000; Kuhn et al. 2000; Wilhelm and Beishuizen 2003; Zimmerman 2007). In diesem Prozess stellt evidenzbasiertes Argumentieren, d. h. das Begründen gezogener Schlussfolgerungen anhand experimenteller Daten, ein wichtiges Mittel dar, um gefundene Ergebnisse sachgemäß zu interpretieren und Erkenntnisse nachvollziehbar zu kommunizieren (Budke and Meyer 2015). Die Fähigkeit, Interpretationen wissenschaftlicher Untersuchungen mithilfe vorliegender Daten zu begründen und damit evidenzbasiert zu erläutern, gehört zu den domänenübergreifenden Aspekten des wissenschaftlichen Denkens (Klahr 2000; Kuhn et al. 2000). In der vorliegenden Arbeit wird evidenzbasiertes Argumentieren im Kontext experimenteller Hypothesentestung als eine wesentliche Facette wissenschaftlichen Denkens bei Schüler*innen der 5. und 6. Klassenstufe untersucht.

1 Theoretischer Hintergrund

1.1 Experimentelle Hypothesentestung und evidenzbasiertes Argumentieren

Experimentelle Hypothesentestung dient der Überprüfung von Kausalhypothesen, und damit letztlich der Entwicklung von wissenschaftlichen Kausalmodellen und Theorien. Die Erfassung und Förderung der zum hypothesentestenden Experimentieren benötigten Fähigkeiten spielt daher eine wichtige Rolle, wenn im schulischen Kontext wissenschaftliches Denken als domänenübergreifende Facette gefördert werden soll (so z. B. im K-12 Science Education Framework, 2012, sowie im Lehrplan 21 der Schweiz, 2016). Im hier vorliegenden Kontext der experimentellen Hypothesentestung bezeichnet evidenzbasiertes Argumentieren eine wichtige Teilfähigkeit, nämlich die Begründung von Schlussfolgerungen über mögliche Kausalzusammenhänge (als Argument) anhand von experimentell gefundenen Daten (Evidenz). Diese Definition entspricht den Konzeptualisierungen des Argumentierens im naturwissenschaftlichen Unterricht als Fähigkeit, Behauptungen zu begründen und zu rechtfertigen (Sampson and Clark 2008). In unserem Gebrauch des Begriffs orientieren wir uns insbesondere an den Arbeiten Deanna Kuhns, die als Kern wissenschaftlichen Denkens und Argumentierens die Fähigkeit ansieht, ein mentales Modell der zu untersuchenden kausalen Zusammenhänge aufzubauen. In diesem Modell werden Theorie (Hypothesen) und Daten (Evidenz) a) zuverlässig voneinander unterschieden und b) aufeinander in einer Weise bezogen, die die nachvollziehbare und widerspruchsfreie Überprüfung aufgestellter Hypothesen ermöglicht (Kuhn 2000, 2011). Dieser in erster Linie innere Prozess ist eine wichtige Grundlage, um die so gebildeten Argumente im Sinne kollaborativen Argumentierens auch mit anderen austauschen zu können (Kuhn 2000; Osborne 2010). Letztlich dient evidenzbasiertes Argumentieren, als Facette des breiter gefassten Begriffs des wissenschaftlichen Argumentierens, dem persönlichen und kollektiven Wissensaufbau in Bezug auf Fragen nach Sinn auf der Welt und der Erklärung ihrer Phänomene (Driver et al. 2000).

Für die Koordination von Theorie und Evidenz (Kuhn 2000, 2011) durch evidenzbasierte Argumentation im Rahmen der experimentellen Hypothesentestung ist ein Verständnis des Prinzips der Variablenkontrolle bzw. der umzusetzenden Variablenkontrollstrategie essentiell. Die Variablenkontrollstrategie (im Original *control-of-variables-strategy* (CVS); im Deutschen VKS; Chen and Klahr 1999) beschreibt die Fähigkeit, beim experimentellen Testen einer Kausalhypothese ausschließlich die zu untersuchende (unabhängige) Variable zu verändern und alle anderen (unabhängigen) Variablen konstant zu halten. Mit der korrekten Anwendung der VKS kann ein unkonfundiertes Experiment durchgeführt werden, um Haupteffekte einzelner, unabhängiger Variablen nacheinander zu testen (Chen and Klahr 1999; Dewey 2002; Klahr 2000; Tschirgi 1980). In der früheren entwicklungspsychologischen Forschung wurde davon ausgegangen, dass Kinder vor der frühen Adoleszenz kein Verständnis der VKS entwickeln können (Inhelder and Piaget 1958; Klahr et al. 1993; Siegler et al. 1973; Tschirgi 1980). Neuere Forschungsergebnisse konnten diese Annahme allerdings widerlegen. So zeigte sich, dass bereits jüngere Kinder fähig sind, unkonfundierte von konfundierten Experimenten zu unterscheiden (So-

dian et al. 1991) und bereits im Alter von 8 Jahren unkonfundierte Experimente konfundierten Experimenten vorzuziehen (Bullock et al. 2009; Bullock and Ziegler 1999). Bullock and Ziegler (1999) fanden außerdem, dass mehr als 50 % der Schüler*innen der 4. Klassenstufe, fast 80 % der Schüler*innen der 5. Klassenstufe und fast 100 % der Schüler*innen der 6. Klassenstufe ihre Wahl auch in Einklang mit der VKS begründen. Dennoch scheint es nicht so, als würde sich die Fähigkeit, die VKS korrekt zu verstehen und anzuwenden bei allen Individuen ohne pädagogische Unterstützung von selbst zu entwickeln (Zimmerman 2007; Zimmerman and Croker 2013). Über das Jugendalter hinaus bis ins Erwachsenenalter zeigen Studien bei zahlreichen Personen ein nach wie vor unausgereiftes Verständnis der VKS auf prozeduraler und metakonzeptueller Ebene (Kuhn 2007; Schwichow et al. 2020; Zimmerman and Croker 2013; Zimmerman and Klahr 2018, für einen Überblick).

Konzeptuell (Chen and Klahr 1999) und empirisch (Schwichow et al. 2016) lassen sich vier unterschiedlich schwierige Facetten der VKS zeigen: (1) Das Identifizieren eines vorgegebenen Vergleichs als konfundiert oder unkonfundiert (Identifizierung); (2) das selbständige Planen eines unkonfundierten Experiments (Planung); (3) das Interpretieren von Evidenz aus unkonfundierten Experimenten (Interpretation) und (4) die Einsicht in die Tatsache, dass aus konfundierten (d. h. nicht der VKS entsprechend angelegten) Experimenten keine sicheren Schlussfolgerungen bezüglich des kausalen Einflusses einer Variable gezogen werden können (Verständnis). Schwichow et al. (2020) konnten zeigen, dass die letztgenannte Facette die schwierigste ist.

1.2 Einflussfaktoren auf die Qualität evidenzbasierter Argumentation

Evidenzbasierte Argumentation – im Sinn des Ziehens und Begründens von Schlussfolgerungen aus experimenteller Evidenz – wird in der Regel im Zusammenhang mit der Interpretation von im Rahmen der experimentellen Hypothesenprüfung (ggf. durch die Schüler*innen selber) gewonnenen Daten untersucht. So gibt es einige Studien zum gegenseitigen Einfluss von Lernen durch Experimentieren (inquiry-based learning) und evidenzbasiertem Argumentieren (siehe unter anderem Keselman 2003; Kuhn 2000; Kuhn et al. 2008, 2009). Dabei wird vor allem getestet, ob Kinder fähig sind, eindeutige Schlussfolgerungen aus empirischen Daten zu ziehen. Die dafür beschriebenen Datengrundlagen enthalten typischerweise mehrere unabhängige Variablen, deren kausaler Einfluss auf eine oder mehrere abhängige Variablen getestet wird. Die Kinder werden mit den Ergebnissen aus experimentellen Vergleichen konfrontiert und gebeten, die Bedeutung dieser Ergebnisse in Bezug auf eine bestimmte Kausalhypothese zu interpretieren (Interpretation) und diese Schlussfolgerung zu begründen (Argumentation). Wünschenswerte Antworten von Schüler*innen in dieser Art von Aufgabenstellung beinhalten sowohl eine korrekte Interpretation (im Sinne der VKS) als auch eine Argumentation, die auf die empirischen Ergebnisse (die Evidenz) in adäquater Weise Bezug nimmt.

Keselman (2003) kam in ihrer Studie zu dem Schluss, dass ein verbessertes Verständnis der VKS auch zu einer Verbesserung der gebildeten mentalen Kausalmodelle (Kuhn 2000, 2002) führt, und damit auch zu einer Erhöhung der Qualität evidenzbasierter Argumentation. Der Grad an Verständnis der VKS lässt sich also

durch die Qualität von evidenzbasierten Argumentationen abbilden und analysieren (Ryu and Sandoval 2012). Auch Kuhn (2002) nimmt an, dass explizites Wissen über die VKS anhand von Begründungen der Schüler*innen betrachtet werden kann.

Osterhaus et al. (2015) beschreiben drei qualitative Ebenen des evidenzbasierten Argumentierens beim Experimentieren für Schüler*innen der 3. und 4. Klassenstufe: Eine naive Ebene, eine intermediäre Ebene und eine fortgeschrittene Ebene. Unter der naiven Ebene verstehen die Autor*innen Begründungen, die auf den produzierten Effekt (das Ergebnis des Experiments) hinweisen, ohne eine weitere Erklärung abzugeben. Die intermediäre Ebene beschreibt eine Zwischenstufe, die auf nicht relevante experimentelle Gegebenheiten hinweist oder auf den Mechanismus, der zum Ergebnis des Experiments geführt hat. Die fortgeschrittenste Ebene bildet schließlich eine datenbasierte Begründung, die im unkonfundierten Fall anhand der gefundenen Daten die gezogene Schlussfolgerung erläutert und im konfundierten Fall auf den Fehler im experimentellen Design hinweist. Die Qualität der evidenzbasierten Argumentation von Schüler*innen wird von verschiedenen Faktoren beeinflusst. Zunächst steigt mit dem Alter der Schüler*innen auch ihre grundsätzliche Argumentierkompetenz; dies gilt auch für die evidenzbasierte Argumentation: Haslbeck (2019) zeigte einen Zuwachs der Qualität der Argumente von Schüler*innen von der 3. zur 4. Klassenstufe und Edelsbrunner (2017) konnte über die Klassenstufen 1–6 hinweg eine Zunahme der Qualität evidenzbasierter Argumentation feststellen, vor allem über die 3. bis 5. Klassenstufe hinweg.

Die Fähigkeit, anhand empirischer Daten nachvollziehbar zu argumentieren, hängt außerdem vom Aufgabenkontext und dabei insbesondere vom konzeptuellen Vorwissen (bzw. dem domänenspezifischen Inhaltswissen) der Schüler*innen ab (Koslowski 2012). Diverse Studien schreiben dem Vorwissen von Kindern beim evidenzbasierten Argumentieren eine zentrale Rolle zu (Croker and Buchanan 2011; Kuhn et al. 1988; Lazonder and Harmsen 2016; Lazonder et al. 2008; Wilhelm and Beishuizen 2003). McNeill (2011) und Koslowski (2012) kamen in ihren Studien zum Schluss, dass die Fähigkeit, evidenzbasiert zu argumentieren (also Schlussfolgerungen anhand der vorgegebenen Daten zu begründen), direkt vom Grad des konzeptuellen Vorwissens und des Aufgabenkontextes abhängt. Lazonder et al. (2008) fanden, dass Schüler*innen eher vorwissensbasiert argumentierten, wenn sie mit den Inhalten der erfragten Kontexte bereits konzeptuell vertraut waren. Ähnliche Ergebnisse berichteten auch Chinn and Malhotra (2002), die anhand von vier Experimenten untersuchten, ob Schüler*innen der 4.–6. Klassenstufe fähig sind, ihre ursprüngliche Vorstellung anzupassen, nachdem sie empirische Daten interpretieren sollten. Die Autoren fanden heraus, dass die Kinder Mühe damit hatten, sobald die Ergebnisse ihrem Vorwissen widersprachen. Obwohl sich viele Kinder in ihren Argumenten eher auf plausible Hypothesen stützten (Klahr et al. 1993), erkennen sie durchaus sinnvolle und korrekte Vergleiche, sofern diese mit ihren ursprünglichen Vorstellungen übereinstimmen (Croker and Buchanan 2011; Gopnik and Schulz 2004; Sodian et al. 1991). Dies gilt auch andersherum: Kinder bilden Hypothesen, die ihren ursprünglichen Vorstellungen entsprechen (Croker and Buchanan 2011; Kuhn et al. 1995; Schauble 1996). Wenn die Ergebnisse nicht mit ihren ursprünglichen Ansichten übereinstimmen, tendieren Kinder dazu, diese zu ignorieren oder zu verzerren. Ebenso interpretieren Kinder die vorliegenden Daten

häufig fälschlicherweise so, dass ihre ursprünglichen Ansichten damit erklärt werden können (Chinn and Brewer 1993; Chinn and Malhotra 2002; Masnick and Klahr 2003), und es zeigt sich eine generelle Tendenz bei Kindern, Ergebnisse anhand ihres Vorwissens zu interpretieren und damit zu begründen (Lazonder et al. 2008). Einerseits erfordert wissenschaftliches Argumentieren konzeptuelles Vorwissen, um angemessene Begründungen konstruieren und experimentelle Daten einordnen zu können (Osborne et al. 2004, 2016). Andererseits kann das Vorhandensein unhinterfragter Überzeugungen und (Alltags-)Theorien in einem Inhaltsbereich dazu führen, dass Kinder sich in ihren Schlussfolgerungen und ihrer Argumentation auch im Kontext experimenteller Hypothesenprüfung eher von ihrem Vorwissen leiten lassen und vorgefundene Evidenz ignorieren (siehe Chinn and Brewer 1993; Chinn and Malhotra 2002; Kuhn et al. 1988; Masnick et al. 2002).

Zusätzlich zur Vorwissensnähe beeinflusst schließlich auch die Komplexität der zu interpretierenden Daten die Qualität evidenzbasierter Argumentationen. Schüler*innen zeigen in komplex strukturierten Kontexten größere Schwierigkeiten beim evidenzbasierten Argumentieren als in einfacher strukturierten Kontexten (McNeill 2011). Bei der Interpretation von komplexeren Datenlagen haben auch Jugendliche (Kelly and Takao 2002) und sogar Erwachsene (Kruglanski and Gigerenzer 2011) unter Umständen Probleme, gute evidenzbasierte Argumente zu formulieren.

1.3 Interventionen zur Förderung der Qualität evidenzbasierter Argumentation

Obwohl evidenzbasiertes Argumentieren als ein zentraler Bestandteil der Wissenserweiterung durch wissenschaftliches Experimentieren angesehen wird (Lunetta et al. 2007), nutzen Schüler*innen Experimente selten spontan als Argumentationsgelegenheit (Kind et al. 2011; Ludwig 2017; Rod Watson et al. 2004). McNeill (2011) untersuchte die strukturelle Qualität evidenzbasierter Argumentationen bei Schüler*innen der 5. Klassenstufe und identifizierte dabei mehrere Defizite. Die Qualität der von ihm beobachteten Argumentationen verbesserte sich nur mit ausreichender Unterstützung und Instruktion über einen längeren Zeitraum hinweg. Kind et al. (2011) fanden, dass Schüler*innen bezüglich der Diskussion von Experimenten am besten abschnitten, wenn sie nicht selbst experimentiert hatten. Die Autor*innen schlussfolgerten, dass eigenständiges Experimentieren nicht automatisch evidenzbasiertes Argumentieren fördert. Sie kamen zum Schluss, dass dazu Unterstützung notwendig ist, beispielsweise in Form von strukturierten Hilfestellungen bei der Interpretation der experimentellen Daten (Kim and Song 2006; Kind et al. 2011). Sowohl ein Verständnis der VKS als auch die Fähigkeit zum evidenzbasierten Argumentieren lassen sich durch gezielte Interventionen in der Tat fördern. Chinn and Malhotra (2002) gaben Schüler*innen der 4.–6. Klassenstufe gezielte Instruktionen zum evidenzbasierten Argumentieren bei Schlussfolgerungen, die nicht mit den ursprünglichen Vorstellungen übereinstimmten, was zu einer Steigerung der Qualität der evidenzbasierten Begründungen der Kinder führte. Peteranderl (2019) zeigte, dass evidenzbasiertes Argumentieren bei Schüler*innen der 5. und 6. Klassenstufe durch ein explizites Training aller vier Facetten der VKS positiv beeinflusst werden kann: Kinder, die ein VKS-Training erhalten hatten (Experimentalgruppe), begrün-

deten ihre Aussagen zu experimentell gezeigten Haupteffekten anschließend häufiger anhand der vorgefundenen Daten als Kinder in einer nicht in der VKS trainierten Kontrollgruppe.

1.4 Evidenzbasiertes Schlussfolgern und evidenzbasierte Argumentation im Haupteffekte- vs. Interaktionseffekte-Kontext

Experimentelle Hypothesentestung und evidenzbasiertes Argumentieren von Schüler*innen werden in der Regel im Kontext von sehr einfach strukturierten Kausalsystemen getestet, in denen einzelne unabhängige Variablen konsistente und voneinander unabhängige Effekte auf eine abhängige Variable aufweisen. Die Wahl solcher einfach strukturierten Kausalsysteme liegt nahe, wenn man davon ausgeht, dass Kinder zunächst Schwierigkeiten haben, überhaupt Theorie und Evidenz voneinander zu unterscheiden und sich auf das Prinzip der evidenzbasierten, experimentellen Hypothesenprüfung einzulassen. Gerade die Anfänge der Forschung zur Entwicklung von Fähigkeiten und Kompetenzen beim Experimentieren bei Kindern deuteten auf substantielle Schwächen in deren Verständnis der Zusammenhänge von Daten, kausalen Hypothesen und erklärenden Theorien hin (Klahr 2000; Klahr et al. 1993; Kuhn et al. 1988, 1995). Kuhn et al. (2000) erklärt dies mit der Annahme, dass Kinder teilweise inkorrekte mentale Modelle kausaler Zusammenhänge mitbringen, die nicht leicht veränderbar sind (Kuhn et al. 2000; Kuhn 2002). Bevor Kinder beginnen können, ihre mentalen Kausalmodelle anhand vorgefundener Daten anzupassen, ist nach Kuhn ein Verständnis des Prinzips konsistenter und additiver Effekte auf ein sich daraus kumulativ ergebendes Ergebnis zwingend erforderlich. Das Kind muss sich zusätzlich darauf verlassen, dass verschiedene Variablen das Ergebnis unabhängig voneinander und konsistent beeinflussen. Dieses sogenannte „korrekte Analysemodell“ sieht Kuhn als eine Bedingung dafür, dass die Variablenkontrollstrategie korrekt ausgeführt und zur Begründung von Schlussfolgerungen anhand gefundener Daten herangezogen werden kann.

Das „korrekte Analysemodell“ Kuhns, mit additiven, konsistenten und voneinander unabhängigen Effekten entspricht in der statistischen (varianzanalytischen) Sichtweise einem Analysekontext, in welchem nur Haupteffekte eine Rolle spielen. Wir nennen den klassischerweise zur Untersuchung und Förderung hypothesentestenden Experimentierens und evidenzbasierten Argumentierens eingesetzten Kontext daher hier den „Haupteffekte”-Kontext. In der Realität sind die zu beschreibenden Kausalstrukturen in der Regel komplexer. Formale Systematisierungen von Kausalmodellen unterschiedlicher Komplexität finden sich beispielsweise in kognitionspsychologischen Arbeiten zum schlussfolgernden Denken (Rottman et al. 2012), oder als Grundlage der Entwicklung von Tests zum komplexen Problemlösen (Greiff et al. 2012).

Keselman (2003) erweiterte die Komplexität des von ihr herangezogenen Kausalmodells dahingehend, dass nicht, wie sonst oft üblich, nur der Einfluss einer unabhängigen Variable, sondern die (allerdings immer noch additiven) Effekte mehrerer unabhängiger Variablen eine Rolle spielten, welche gleichzeitig auf eine abhängige Variable wirkten (multivariable Kausalität). Eine Interventionsstudie mit Schüler*innen der 6. Klassenstufe in einer Kontroll- und zwei Experimentalgruppen

zeigte, dass explizite Instruktion zur Hypothesenbildung im Kontext multivariabler Kausalzusammenhänge zu einer Verbesserung des hypothesentestenden Experimentierens und des evidenzbasierten Argumentierens führte (Keselman 2003).

Experimentelle Hypothesentestung und evidenzbasierte Argumentation im Rahmen eines über einen additiven „Haupteffekte"-Kontext hinausgehenden Kontexts wurden unseres Wissens nach bei Kindern im Grundschulalter bisher noch nicht systematisch untersucht. Besonders interessant scheint uns die Erweiterung auf einen „Interaktionseffekte"-Kontext. Der Kontext einer Interaktion liegt dann vor, wenn Variablen nicht nur additiv auf ein Ergebnis wirken, sondern sich in ihren Auswirkungen gegenseitig beeinflussen (interagieren). Die Exploration von Interaktionseffekten verlangt nicht nur die systematische Testung der Effekte einzelner unabhängiger Variablen, sondern auch die Beobachtung der Veränderung des Einflusses einer Variable, wenn die andere, interagierende Variable, unter verschiedenen Bedingungen mehrfach getestet wird. Obwohl ein Verständnis der VKS dafür weiterhin wichtig bleibt, ist es allein nicht mehr ausreichend. Das Prinzip evidenzbasierten Argumentierens jedoch, also die Anforderung, gezogene Schlussfolgerungen anhand der gegebenen Daten zu begründen, behält auch in einem „Interaktionseffekte"-Kontext seine Wichtigkeit. In der vorliegenden Studie wird daher gezielt das evidenzbasierte Argumentieren anhand vorgegebener experimenteller Evidenz in einem „Interaktionseffekte"-Kontext untersucht.

1.5 Fragestellung der vorliegenden Studie

In der vorliegenden Arbeit untersuchten wir in einem vorwiegend explorativen Design, in welchem Ausmaß Schüler*innen der 5. und 6. Klassenstufe sich im Rahmen einer vorstrukturierten Argumentationsaufgabe in ihrer Argumentation (der Begründungen ihrer Schlussfolgerungen) auf vorgegebene Evidenz beziehen. Anders als viele bisherige Untersuchungen zur Argumentation mit Hilfe der Variablenkontrollstrategie betrachteten wir den Fall eines Experimentes mit zwei Faktoren, die in ihren Auswirkungen interagieren („Interaktionseffekte"-Kontext). Zielgruppe der Untersuchung waren Schüler*innen der 5. und 6. Klassenstufe an Grundschulen der Deutschschweiz (die Grundschule umfasst dort die Klassen 1–6).

In einem ersten Schritt untersuchten wir die Frage, in welchem Ausmaß Schüler*innen der 5. und 6. Klassenstufe vorgegebene experimentelle Evidenz heranziehen, um Kausaleffekte in einem „Interaktionseffekte"-Kontext zu interpretieren bzw. ihre Interpretation argumentativ zu begründen.

> Fragestellung (1): *In welchem Ausmaß nutzen Schüler*innen der 5. und 6. Klassenstufe vorgegebene experimentelle Evidenz in ihrer Argumentation in einem „Interaktionseffekte"-Kontext?*

Zur Klassifizierung der Antworten der Schüler*innen wurde ein Kodierschema entwickelt, um unterschiedliche Grade an Evidenzbasierung zu differenzieren. Die Häufigkeiten der mit Hilfe dieses Schemas kodierten schriftlichen Antworten der Schüler*innen (eine kategoriale, ordinale Reihung mit aufsteigendem Grad an Evidenzbasierung) bildeten die Grundlage statistischer Analysen. Dabei interessierten uns insbesondere Effekte der Klassenstufe sowie des Aufgabentyps auf die Qualität

(im Sinn von Evidenzbasierung) der Argumentation in den schriftlichen Antworten. In Bezug auf Effekte der Klassenstufe existieren bereits empirische Befunde, welche eine Zunahme der Argumentationskompetenz im Kindesalter zeigen (Edelsbrunner 2017; Osterhaus et al. 2015; Haslbeck 2019); daher nahmen wir an, dass ältere Schüler*innen (6. Klassenstufe) verglichen mit jüngeren Schüler*innen (Klassenstufe 5) ein höheres Ausmaß an evidenzbasierter Argumentation zeigen würden. In Bezug auf mögliche Effekte des Aufgabentyps waren unsere Analysen explorativer Natur. Die zwei uns zur Verfügung stehenden Aufgaben unterschieden sich sowohl in der Vorwissensnähe, als auch in der Art der statistischen Interaktion (ordinale vs. disordinale Interaktion) und somit der Komplexität des heranzuziehenden Kausalmodells. Eine ausführliche Erläuterung zu den Aufgabentypen wird in Sect. 2.4.2 gegeben.

Die zweite Fragestellung bezog sich auf die Effekte eines expliziten VKS-Trainings, welches innerhalb eines experimentellen Designs mit Experimental- und Kontrollgruppe realisiert wurde. Inhalt des Trainings waren alle vier der nach Chen and Klahr (1999) und Schwichow et al. (2016) unterscheidbaren Aspekte der Variablenkontrollstrategie: Das Planen und Identifizieren von unkonfundierten Experimenten, das Interpretieren von Evidenz aus unkonfundierten Experimentieren und das Begründen der gezogenen (oder nicht gezogenen) Schlussfolgerungen anhand des Prinzips eines „fairen“, d.h. unkonfundierten Vergleichs. Eine Bestätigung der generellen Wirksamkeit dieses VKS-Trainings auf die Anwendung der VKS und auf die Qualität evidenzbasierten Argumentierens in einem „Haupteffekte“-Kontext liegt bereits vor (Peteranderl 2019). In der hier vorgestellten Studie soll anhand einer Reanalyse vorhandener Daten aus der Studie von Peteranderl (2019) darüber hinaus festgestellt werden, ob sich das Training zur Variablenkontrollstrategie auch auf evidenzbasierte Argumentation bei der Interpretation experimenteller Evidenz in einem „Interaktionseffekte“-Kontext auswirkt.

Fragestellung (2): *Zeigt ein VKS-Training zur Anwendung der Variablenkontrollstrategie im „Haupteffekte“-Kontext Auswirkungen auf die Fähigkeit zur evidenzbasierten Argumentation bei der Interpretation experimenteller Evidenz in einem „Interaktionseffekte“-Kontext?*

2 Methode

2.1 Stichprobe und Design

Die Stichprobe bestand aus 29 fünften und 9 sechsten Klassen (insgesamt 38 Klassen) aus Grundschulen in der Deutschschweiz. Die Datenerhebung der vorliegenden Studie erfolgte im Rahmen der Schweizer MINT-Studie (siehe beispielsweise Schalk et al. 2019), in welcher die Effekte der Einführung physikalischer Konzepte im Sachunterricht sowie des expliziten Trainings der Variablenkontrollstrategie auf die Experimentierkompetenz von Schüler*innen untersucht wurden. Insgesamt bestand die Stichprobe aus $n = 758$ Schüler*innen. Von diesen wurden $n = 140$ Schüler*innen für die Analysen ausgeschlossen, davon $n = 62$, die am Training nicht vollständig teilgenommen hatten, $n = 48$ mit fehlenden Einverständniserklärungen zur Ver-

arbeitung ihrer Daten, $n = 11$, die an sämtlichen Erhebungs- und Trainingstagen abwesend waren sowie $n = 19$, die am Vor-, jedoch nicht am Nachtest anwesend waren. Die Datenbasis umfasste demnach eine Gesamtstichprobe von $N = 618$ Schüler*innen ($\text{Mittelwert}_{\text{Alter}} = 11.67$, $\text{SD}_{\text{Alter}} = 0.65$, 309 Mädchen, 309 Jungen). In allen Klassen wurden sowohl eine Experimentalbedingung (VKS-Training) als auch eine Kontrollbedingung (Kontrolltraining) realisiert. Die Schüler*innen wurden randomisiert innerhalb der Klasse entweder der Experimentalgruppe ($n = 318$) oder der Kontrollgruppe ($n = 300$) zugeordnet.

2.2 Ablauf der Studie

Wie Abb. 1 zeigt, begann die Datenerhebung für alle Kinder mit dem Ausfüllen eines Vortests am ersten Messzeitpunkt, etwa eine Woche vor Beginn der Trainingsintervention. Im Vortest (VT) bearbeiteten die Schüler*innen einen Experimentierfähigkeitstest, der im Rahmen des übergeordneten Forschungsprojektes entwickelt und ausführlich psychometrisch evaluiert wurde (für detaillierte Informationen siehe Peteranderl 2019). Zusätzlich wurden zu diesem ersten Messzeitpunkt mehrere Kovariaten (unter anderem kognitive Fähigkeiten und Leseverständnis) erhoben, die für die Fragestellung der vorliegenden Studie nicht relevant sind und daher nicht berücksichtigt werden. Anschließend wurden die Schüler*innen in sämtlichen Klassen per Zufall in zwei Gruppen geteilt, von denen jeweils eine Gruppe das VKS-Training (Experimentalbedingung) und die andere Gruppe das Kontrolltraining (Kontrollbedingung) durchlief. Die Trainings liefen zeitgleich in unterschiedlichen Räumen ab und die Gruppen wechselten anschließend nicht. Beide Trainings umfassten je drei Lektionen über einen Zeitraum von zwei Wochen. Etwa eine Woche nach den abgeschlossenen Trainings fand der zweite Messzeitpunkt statt, an dem die gesamte Klasse den Experimentierfähigkeitstest als Nachtest (NT) ausfüllte.

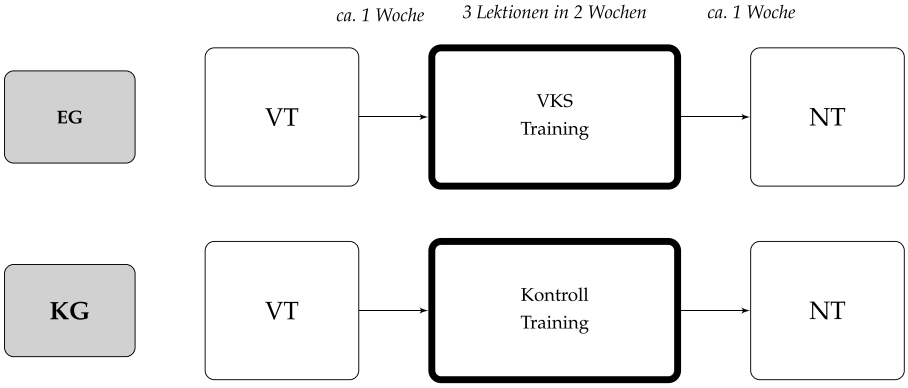


Abb. 1 Überblick über den Ablauf der Testungen. *KG* = Kontrollgruppe, *EG* = Experimentalgruppe, *VT* = Vortest (1. Messzeitpunkt), *NT* = Nachtest (2. Messzeitpunkt)

2.3 Trainingsinterventionen

2.3.1 Experimentalgruppe: Variablenkontrollstrategie (VKS) Training

Ziel des VKS-Trainings war das Erlernen und korrekte Anwenden aller vier Facetten der Variablenkontrollstrategie nach Schwichow et al. (2016). Das VKS-Training umfasste mehrere Demonstrationsexperimente, eine explizite Einführung in die Variablenkontrollstrategie sowie das gezielte Lernen und Anwenden der einzelnen Facetten der VKS. In der ersten Trainingslektion wurden den Kindern mehrere konfundierte und unkonfundierte Demonstrationsexperimente mit unterschiedlichen Materialien gezeigt. Konfundierte Experimente wurden gezielt genutzt, um kognitive Konflikte bei den Schüler*innen hervorzurufen (z. B. durch das Sammeln von Interpretationsvorschlägen und der Frage zur fälschlich gezogenen Schlussfolgerung: „Können wir uns da ganz sicher sein?“). Anhand von unkonfundierten Experimenten wurde die Variablenkontrollstrategie veranschaulicht und detailliert erklärt. Anschließend mussten die Schüler*innen in Paaren vorbereitete Experimente als „fair“ (unkonfundiert) oder „unfair“ (konfundiert) erkennen und jeweils begründen, ob und weshalb man eine Schlussfolgerung ziehen konnte oder nicht. Während der zweiten Trainingslektion, etwa eine Woche später, wurden die Schüler*innen darin angeleitet, sich in Kleingruppen mit speziell für das Einüben der VKS konstruierten Kugelbahnen (vergleichbar mit den Rampen aus Chen and Klahr 1999) auseinanderzusetzen und kontrollierte Experimente zu selbst gewählten abhängigen Variablen zu planen, diese durchzuführen, zu interpretieren und alles auf einem dafür vorbereiteten Arbeitsblatt zu dokumentieren.

2.3.2 Kontrollgruppe: Training ohne VKS

Auch die Kinder der Kontrollgruppe führten während ihres Unterrichts selbstständig Versuche durch, erhielten jedoch keine Einweisung in das hypothesentestende Experimentieren. Das Ziel des Kontrolltrainings war vielmehr die Einführung von inhaltlichem Wissen über Stromkreise und Schaltskizzen. Die Schüler*innen arbeiteten dazu mit Bauelementen wie Batterien, Kabeln und Klemmen, Schaltern sowie zweifarbig leuchtenden LEDs. Der Unterricht wurde durch die vertrauten Klassenlehrpersonen der Kinder durchgeführt; diese hatten dafür vorab eine ausführliche Einführung in die Materialien und den Ablauf, ein detailliertes Skript für alle Lektionen und dafür entwickelte Arbeitsblätter erhalten. In der ersten Lektion erklärte die Lehrperson den Schüler*innen das Material und demonstrierte die korrekte Handhabung. Anschließend wurden die Schüler*innen von der Klassenlehrperson dabei angeleitet, in Partnerarbeit auf vorstrukturierten Arbeitsblättern zwischen Serien- und Parallelschaltungen zu unterscheiden. In der zweiten Lektion, etwa eine Woche später, sollten die Schüler*innen in Kleingruppen eigene Schaltkreise bauen und zeichnen, um z. B. eine Sicherungsschaltung darzustellen. Während des gesamten Kontrolltrainings fand keinerlei Instruktion bezüglich der Variablenkontrollstrategie statt.

2.4 Experimentierfähigkeitstest

2.4.1 Gesamttest

Der Experimentierfähigkeitstest beinhaltete insgesamt 15 Fragen, die unterschiedliche Aspekte wissenschaftlichen Denkens abdeckten, darunter die vier Facetten der Variablenkontrollstrategie (Schwichow et al. 2016) sowie typische Fehlkonzepte bei der Planung von Experimenten (Siler and Klahr 2012; Schauble et al. 1991). Eine detaillierte Beschreibung aller Fragetypen findet sich in Peteranderl (2019). Der Kern unserer Analysen befasst sich mit den Unteraufgaben zum evidenzbasierten Argumentieren im „Interaktionseffekte"-Kontext (Fragetyp „Interaktion"), welche ebenfalls ein Bestandteil des Tests waren.

2.4.2 Fragen zum „Interaktionseffekte" -Kontext

Für die vorliegende Studie sind ausschließlich die Aufgaben vom Fragetyp „Interaktion" relevant. Dieser Fragetyp besteht aus insgesamt vier Fragen, welche in zwei Aufgaben eingebettet sind. Beide Aufgaben beschreiben anhand einer kurzen einleitenden Geschichte eine experimentelle Situation mit zwei unabhängigen Variablen mit je zwei Ausprägungen und einer abhängigen Variablen. Anschließend werden den Schüler*innen jeweils vier Bilder präsentiert, welche vier mögliche Kombinationen aus beiden unabhängigen Variablen zeigen. Unterhalb jedes Bildes sind die jeweiligen Ausprägungen der beiden unabhängigen Variablen und das dadurch erreichte Ergebnis (die Ausprägung der abhängigen Variablen) angegeben. Die Aufgabe der Schüler*innen besteht darin, für jede der beiden unabhängigen Variablen separat die Frage zu beantworten, ob diese Variable einen Einfluss auf das Ergebnis hat und warum. Da die vorgegebenen Ergebnisse allen möglichen Kombinationen der unabhängigen Variablen entsprechen, haben die Schüler*innen alle dafür notwendigen Informationen. Das Antwortformat ist in zwei Schritte aufgeteilt: zunächst sollen die Schüler*innen im Single-Choice Format („Hat [diese Variable] einen Einfluss?") eine Antwort ankreuzen („Ja", „Nein", „Das kann man nicht sagen",„Es kommt darauf an"), die sie dann in einem zweiten Schritt („Woher weißt du das?") schriftlich begründen sollen (offenes Antwortformat). Eine Aufgabe („Kühe") ist so konzipiert, dass die Daten einer ordinalen Interaktion entsprechen, die andere („Spieleautomat") so, dass die Daten einer disordinalen (cross-over) Interaktion entsprechen. Bei einer ordinalen Interaktion interagieren die beiden Variablen insofern, dass die Haupteffekte beider Variablen global interpretiert werden können. Bei einer disordinalen Interaktion kann keiner der Haupteffekte global interpretiert werden, es müssen also beide Variablen eine bestimmte Ausprägung zeigen, um das „gewünschte" Ergebnis zu erzielen. Ein weiterer Unterschied betrifft die Vorwissensnähe der für die Aufgabe verwendeten Geschichte. Während die erste Aufgabe (Einflussfaktoren auf die Milchleistung von Kühen) einen starken lebensweltlichen Bezug hat, bei dem die Kinder basierend auf ihrem möglichen Vorwissen Vermutungen über kausale Zusammenhänge aufstellen können, ist die zweite Aufgabe (Effekte von unterschiedlich geformten Hebeln auf die Ausgabe eines fiktiven „Spieleautomaten")

bewusst abstrakt gehalten. Die beiden Aufgaben und jeweils die erste dazugehörige Frage sind in den Abb. 2 und 3 dargestellt.

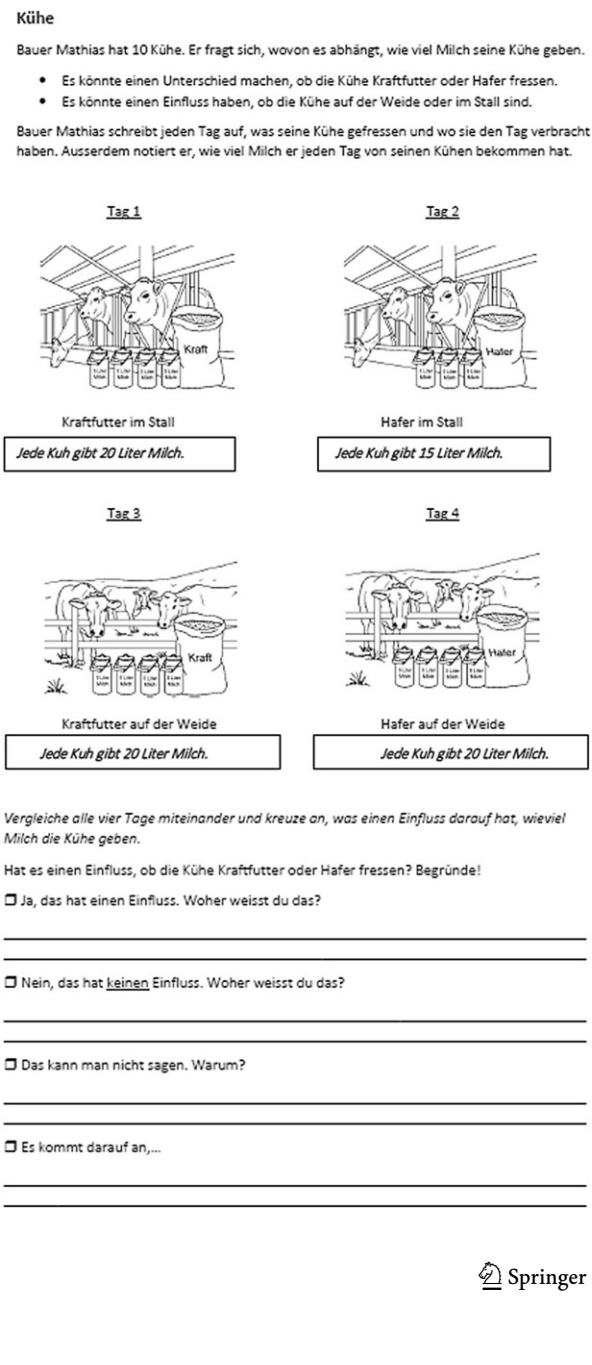


Abb. 2 Kühe-Aufgabe: Aufgabe mit einer ordinalen Interaktion der unabhängigen Variablen und erwartetem Vorwissen der Schüler*innen

Abb. 3 Spieleautomat-Aufgabe: Aufgabe mit disordinaler (cross-over) Interaktion der beiden unabhängigen Variablen und nicht erwartetem Vorwissen der Schüler*innen

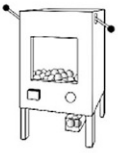
Spieleautomat

Robin versucht zu verstehen, wie der Spielzeug-Kaugummiautomat funktioniert. Er kann

- einen runden Knopf oder einen viereckigen Knopf drücken.
- und anschliessend den rechten oder den linken Hebel betätigen.

Je nachdem, was man tut, wirft der Automat unterschiedlich viele Kaugummi-Kugeln aus. Robin hat diese vier Versuche ausprobiert:

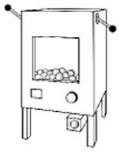
Versuch 1



runder Knopf, rechter Hebel

Robin bekommt 3 Kugeln.

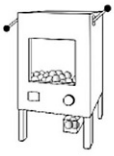
Versuch 2



viereckiger Knopf, rechter Hebel

Robin bekommt 1 Kugel.

Versuch 3



viereckiger Knopf, linker Hebel

Robin bekommt 3 Kugeln.

Versuch 4

runder Knopf, linker Hebel

Robin bekommt 1 Kugel.

Vergleiche alle vier Spieleautomaten miteinander und kreuze an, was einen Einfluss darauf hat, wie viele Kugeln der Spieleautomat auswirft. Begründe!

Hat es einen Einfluss, ob Robin den runden oder den viereckigen Knopf drückt?

❒ Ja, das hat einen Einfluss. Woher weisst du das?

❒ Nein, das hat keinen Einfluss. Woher weisst du das?

❒ Das kann man nicht sagen. Warum?

❒ Es kommt darauf an, ...

3 Kodierschema zur Klassifizierung der argumentativen Evidenzbasierung der schriftlichen Antworten

3.1 Entwicklung des Kodierschemas

Für die vorliegende Studie wurde ein Kodierschema entwickelt, um die Antworten der Schüler*innen nach dem Grad ihrer Evidenzbasierung zu klassifizieren. Dazu wurden einerseits bereits vorliegende Kodierschemata als Orientierungsgrundlage herangezogen, die auf Basis von ähnlichen Aufgaben in einem „Haupteffekte"-Kontext entwickelt wurden (Edelsbrunner 2017; Haslbeck 2019), andererseits wurden die schriftlichen Antworten der Schüler*innen als empirische Datengrundlage zur Entwicklung genutzt. In einem schrittweisen Prozess wurden vorgängig Antworten aus einer Pilotierungsphase des Experimentierfähigkeitstests während dessen Entwicklungsphase kodiert und die Interrater-Reliabilität bestimmt. Durch diese Herangehensweise wurde der theoriebasierte Entwicklungsprozess des Kodierschemas durch einen datenbasierten Entwicklungsprozess ergänzt. Das Resultat war ein Kodierschema mit fünf ordinal gereihten Kategorien, welches in Tab. 1 dargestellt und dessen Anwendung im Folgenden erläutert wird.

Für die Auswertung wurden die schriftlichen Antworten der Schüler*innen auf die vier Fragen der beiden Aufgaben aus dem „Interaktionseffekte"-Kontext herangezogen. Die ausgewählte Antwort im Single-Choice Teil diente dabei lediglich als Hilfsmittel bei der Einordnung der Begründungen der gezogenen Schlussfolgerung im offenen Antwortteil. Sie dienten jedoch nicht zur Selektion der Antworten oder als Entscheidungsgrundlage für die Kategorisierung. Für jede Antwort wurde genau ein Code vergeben. Wurden in einer Antwort mehrere Argumente formuliert, so wurde nur die beste (d. h. am meisten evidenzbasierte) Begründung gewertet.

Jede der vier Fragen fragt nach dem kausalen Einfluss eines spezifischen Faktors. In keinem der vorgegeben Fälle kann die Frage jedoch mit einem Verweis auf einen einfachen Haupteffekt beantwortet (z. B. „Ja, die Art des Futters hat einen Einfluss darauf, wie viel Milch die Kühe geben.") und durch einen entsprechenden Verweis auf die Evidenz aus einem unkonfundierten Vergleich hinreichend begründet werden. Vielmehr muss eine vollständig evidenzbasiert argumentierte Antwort dem Interaktionseffekt Rechnung tragen, indem z. B. auf die Frage nach dem Einfluss des Futters darauf verwiesen wird, dass es darauf ankomme, ob die Kühe das Futter im Stall oder auf der Weide erhielten. Dies gilt auch für den Fall der ordinalen Interaktion zwischen Ort und Futter in der Kühe-Aufgabe: Auch wenn die Milchleistung der Kühe im Mittel auf der Weide höher ist als im Stall (Haupteffekt), so muss der zweite Faktor, das Futter, doch berücksichtigt werden – denn nur, wenn die Kühe Hafer fressen, macht der Ort einen Unterschied. Diese Art der Argumentation setzt allerdings eine Erkenntnis voraus, die nur durch den systematischen Vergleich aller Ausprägungen des vorgegebenen Experimentes erreicht werden kann, also die Einbeziehung der gesamten verfügbaren Evidenz.

In Anlehnung an die Kodiermanuale von Edelsbrunner (2017) und Haslbeck (2019) definieren wir drei Stufen evidenzbasierter Argumentation. Diese unterscheiden sich darin, wie vollständig die Bezüge auf die vorgegebene Datengrundlage (Evidenz) in der Argumentation der Schüler*innen sind. Die fortgeschrittenste Ka-

Tab. 1 Übersicht der fünf Evidenzkategorien mit Bezeichnung, Beispielen und Beschreibung

Bezeichnung	Beschreibung	Beispiele
Unverständliche Argumentation	Sehr kurze Sätze oder **Satzfragmente**, keine Begründung erkennbar **Tautologische** Antworten	*– keine Ahnung* *– Zufall* *– weil ich es weiß*
Vorwissensbasierte Argumentation	Antwort als **Begründung erkennbar** Begründung wird aufgrund des **Vorwissens** des Kindes zum Thema formuliert **Kein** Hinweis auf **evidenzbasierte Schlussfolgerungen**	*– Es kommt darauf an, wie viel sie am Tag getrunken haben* *– Es kommt darauf an, was es für Kühe sind* *– Das kann man nicht sagen, das ist Glückssache*
Ansatzweise evidenzbasierte Argumentation	Eindeutig **falsche, evidenzbasierte Schlussfolgerungen**, gegebenenfalls mit Hinweis auf einen oder mehrere **Vergleiche**, welche/r jedoch **falsch** ist/sind Überlegungen der Kinder sind erkennbar evidenzbasiert, aber **nicht zwingend nachvollziehbar**	*– Ja, das hat einen Einfluss, es steht auf dem Blatt* *– Bei Frage zum Einfluss des Ortes: Es kommt darauf an, weil am Tag 1 gab es 20l mit Kraftfutter und am Tag 2 gab es 15l mit Hafer. [Tag 1 und 2 sind beide im Stall]*
Unvollständig evidenzbasierte Argumentation (Haupteffekt erkannt)	Begründung ist **korrekt evidenzbasiert**: Hinweis auf einen Vergleich mit **korrekter Schlussfolgerung (korrekter Hinweis auf den Haupteffekt)** Hinweis auf **mittleren Wert** beim Kraftfutter oder der Weide (je nach Frage)	*– Frage zum Einfluss des Futters: Ja, das hat einen Einfluss, das kann man bei Versuch Tag 1 und 2 sehen. [Kraftfutter im Stall vs. Hafer im Stall]* *– Es hat einen Einfluss, Kraftfutter gibt mehr Milch*
Vollständig evidenzbasierte Argumentation	Korrekt erkannte Interaktion Korrekter Hinweis auf die Interaktion zwischen den **zwei Variablen** **Futter:** *Hinweis, dass es darauf ankommt, wo die Kühe sich aufhalten (Weide/Stall)* **Ort:** *Hinweis, dass es darauf ankommt, was die Kühe fressen (Hafer/ Kraftfutter)* **Knopf:** *Hinweis, dass es darauf ankommt, welchen Hebel man benutzt* **Hebel:** *Hinweis, dass es darauf ankommt, welchen Knopf man drückt*	*– Frage zum Einfluss des Futters: Es kommt darauf an, ob sie drin oder draussen sind* *– Frage zum Einfluss des Ortes: Es kommt darauf an, was man ihnen zu essen gibt*

tegorie, *vollständig evidenzbasierte Argumentation*, umfasst alle Antworten, in denen die Interaktion zwischen den beiden unabhängigen Variablen korrekt umschrieben wird.

In die Kategorie *unvollständig evidenzbasierte Argumentation* fallen alle Antworten, die korrekte Schlussfolgerungen aus einem Teil der vorgegebenen Daten ziehen (z. B. einen Haupteffekt benennen) und diesen anhand passend ausgewählter Evidenz begründen, dabei jedoch die Interaktion außer Acht lassen. Ein Beispiel wäre die Single-Choice Wahl *Ja, das hat einen Einfluss* zusammen mit der Begründung *Das sieht man an Tag 1 und Tag 2*.

In die Kategorie *ansatzweise evidenzbasierte Argumentation* fallen schließlich alle Versuche einer Begründung, die zwar einen Bezug zu den vorgegebenen Daten herstellen, diese jedoch nicht im Sinn eines gültigen Argumentes nutzen können,

da der Bezug zu unspezifisch ist (*Es steht auf dem Blatt*) oder weil Daten falsch wiedergegeben werden (*Weil sie mit Hafer draußen 20 Liter mehr gibt*).

Nach den drei Kategorien für evidenzbasierte Antworten folgen schließlich noch zwei Kategorien für Antworten, in denen die Kinder sich in keiner Weise auf die vorgegebene Evidenz beziehen. Von besonderem Interesse waren für uns Versuche der Kinder, ihre Auswahl im Single-Choice Teil vorwissensbasiert zu begründen. Unter die Kategorie *vorwissensbasierte Argumentation* fallen Antworten, in denen die Kinder zwar eine inhaltlich sinnvolle Begründung versuchen, dabei jedoch ausschließlich Gründe außerhalb der gegebenen Evidenz ausführen. In der Regel sind diese gestützt auf inhaltliches Vorwissen, bzw. auf persönliche Überzeugungen (z. B. *weil Kraftfutter den Tieren Kraft gibt*).

Zur letzten gebildeten Kategorie, *unverständliche Argumentation*, zählen schließlich Antworten, bei denen weder eine evidenz-, noch eine vorwissensbasierte Begründung erkennbar ist (z. B. *weil ich es weiß*).

3.2 Psychometrische Eigenschaften des Tests und des Kodierschemas

Um die psychometrischen Eigenschaften der Testaufgaben und der anhand des erstellen Kodierschemas kodierten Antworten zu untersuchen, wurden die Interrater-Reliabilität, die Faktorenstruktur, sowie die interne Konsistenz der vier am Vortest sowie am Nachtest anhand des Schemas kodierten Antworten der Schüler*innen untersucht.

Um die Interrater-Reliabilität zu schätzen, wurden etwa 20 % aller kodierten Antworten ($n = 146$) randomisiert ausgewählt und von zwei unabhängigen Beurteilerinnen nochmals kodiert. Für die Berechnung der Interrater-Reliabilität wurde Gwets $AC_1(\gamma)$ berechnet (Gwet 2014), da es ein robusteres Maß als *Cohen's* κ ist (Honda and Ohyama 2020). Tab. 2 zeigt die separaten Interrater-Reliabilitäten und deren Standardfehler über die fünf Evidenzkategorien berechnet für jede Frage. Laut Gwet (2014) sind Werte über 0.20 akzeptabel und die geschätzten Interrater-Reliabilitäten entsprechen guten bis sehr guten Werten.

Anschließend wurde die Faktorenstruktur der Kodierungen am Vor- und Nachtest überprüft. Hierfür wurde eine konfirmatorische Faktorenanalyse geschätzt, in welcher die vier kodierten Antworten am Vortest auf eine latente Variable luden, sowie die vier kodierten Antworten am Nachtest auf eine zweite, korrelierte latente Variable. Es wurde am Vor- und Nachtest jeweils ein Modell mit einer latenten Variablen angenommen, da erwartet wurde, dass evidenzbasierte Argumentation über die unterschiedlichen Aufgaben hinweg eine vergleichbare Fähigkeit darstellt. Für Unterschiede zwischen den beiden Aufgaben wurde vorrangig erwartet, dass sich diese auf Kategoriengrenzen (Wahrscheinlichkeiten, mit denen die einzelnen Kate-

Tab. 2 Interrater-Reliabilität der Kodierung im Vortest, berechnet mit *Gwet's* γ für jede der vier Fragen

Aufgabe	*Gwet's* γ	Standardfehler
Kühe Frage 1	0.77	0.039
Kühe Frage 2	0.74	0.042
Spieleautomat Frage 1	0.82	0.035
Spieleautomat Frage 2	0.80	0.037

gorien auftreten) auswirken. In diesem Modell wurden korrelierte Messfehler derselben Antworten über die Zeit zugelassen. Zusätzlich wurden korrelierte Messfehler der zwei Antworten, welche zu jeder der zwei Aufgaben gehörten, modelliert (vgl. Studhalter et al. 2021). Das Modell wurde mittels *weighted least squares means and variances*-Schätzer, welcher für die kategoriale Datenstruktur angemessen ist, im *R*-Paket *lavaan* (Rosseel 2012) angepasst. Dieses Modell wies eine gute Passung mit den Daten auf, $\chi^2(11) = 31.27$, $p = 0.001$, RMSEA $= 0.06$, CFI $= 1.00$, SRMR $= 0.03$. Die geschätzten Faktorladungen der vier kodierten Antworten waren an jedem Messzeitpunkt signifikant mit $p < 0.001$ und lagen standardisiert recht homogen und hoch im Bereich 0.65–0.74.

Schlussendlich wurde als Schätzer der internen Konsistenz Omega, welches grundsätzlich genauer ist als Cronbach's Alpha (Dunn et al. 2014), im R-Paket *psych* (Revelle 2013) anhand polychorischer Korrelationen, welche für die kategoriale Datenstruktur angemessen sind, geschätzt. Sowohl am Vortest als auch am Nachtest wurde Omega mit 0.84 geschätzt, was einer recht hohen internen Konsistenz entspricht (Feißt et al. 2019).

3.3 Statistische Analysen

Um die erste Fragestellung zu untersuchen, welche sich mit dem Ausmaß beschäftigt, in dem Schüler*innen der 5. und 6. Klassenstufe in ihren Argumentationen auf Evidenz zurückgreifen, werden im Folgenden zunächst deskriptive Analysen über die Häufigkeiten der Argumentationen in den fünf Kategorien des Kodierschemas am Vortest berichtet. Diese deskriptiven Analysen werden erst über beide Klassenstufen und Aufgaben hinweg und danach einzeln für die beiden Klassenstufen sowie für die beiden Aufgaben berichtet. Dann wird zur inferenzstatistischen Absicherung anhand von chi^2-Tests untersucht, ob sich der Anteil evidenzbasierter Argumentationen (d. h. Argumentationen, die mindestens der Stufe *ansatzweise evidenzbasiert* des Kodierschemas zuzuordnen sind) zwischen den beiden Klassenstufen sowie zwischen den beiden Aufgaben unterscheidet.

Um die zweite Fragestellung zu untersuchen, welche sich mit dem Einfluss des VKS-Trainings auf die Argumentation der Schüler*innen beschäftigt, werden erst deskriptive Statistiken berechnet, welche die Anteile der Argumentation in den fünf Kategorien des Kodierschemas am Vor- und Nachtest in der Kontroll- und Experimentalgruppe zeigen. Danach wird als inferenzstatistische Methode eine ordinale Regressionsanalyse geschätzt. Der ordinale Ansatz wurde gewählt, da nicht davon auszugehen ist, dass die fünf Antwortkategorien einer intervallskalierten Variablen entsprechen. Zudem konnte so untersucht werden, ob sich Effekte der Intervention für spezifische Kategorien zeigen. Im VKS-Training wurde geübt, die Variablenkontrollstrategie anzuwenden, um Haupteffekte zu untersuchen. Dies könnte beispielsweise dazu führen, dass sich vor allem die Häufigkeit von Argumentationen, die der Kategorie *unvollständig evidenzbasiert (Haupteffekt erkannt)* zuzuordnen sind, nach dem Training erhöht.

Um inferenzstatistisch zu untersuchen, welche Effekte auf das Argumentieren sich aus den deskriptiven Mustern ableiten lassen und dabei auch Abhängigkeiten innerhalb einzelner Schulklassen zu berücksichtigen, schätzen wir eine ordinale

Mehrebenen-Regressionsanalyse. Dabei wird als abhängige Variable die Kategorie betrachtet, in welche die Argumentationen der Schüler*innen fallen. Eine Modellierung, welche das ordinale Skalenniveau der abhängigen Variablen berücksichtigt, wird mittels kumulativer Logit-Link-Modellierung erreicht (siehe Christensen 2015). Dabei wird die Dichte der ordinalen Antwortkategorien anhand einer latenten Variable mit Mittelwert 0 und (unter Anwendung des Logit-Link) Standardabweichung $pi^2/3$ abgebildet (für eine konzeptuelle Erläuterung dieses Ansatzes siehe Bürkner and Vuorre 2019). Die ordinalen Antwortkategorien summieren sich entlang der Dichteverteilung dieser latenten Variablen auf, bis mit der fünften Kategorie die vollständige Verteilung der Antworten abgebildet ist. Die Dichte der zugrundeliegenden latenten Variable, auf Basis derer die Kategoriengrenzen geschätzt werden, ist in Abb. 6 dargestellt. Welchen Bereich die fünf Kategorien auf der latenten Variable annehmen (d. h., mit welcher Wahrscheinlichkeit Schüler*innen Argumentationen zeigen, welche in die spezifischen Kategorien fallen), wird anhand von vier Kategoriengrenzen geschätzt. Die vier Kategoriengrenzen setzen die Positionen der fünf Kategorien auf der latenten Variablen fest und können mittels Exponentialfunktion in einfach zu interpretierende Wahrscheinlichkeiten dafür, die fünf Kategorien zu zeigen, umgerechnet werden. Die Effekte der unabhängigen Variablen wirken sich additiv (bzw. jene von Interaktionstermen interaktiv) auf die Positionen der vier Kategoriengrenzen aus. Zur weiteren Erläuterung dieses Ansatzes empfehlen wir interessierten Leser*innen Christensen (2018), sowie Bürkner and Vuorre (2019) für eine Perspektive unter Bayesianischer Implementierung.

Als Effekte unabhängiger Variablen wurden Haupteffekte der Bedingung (Experimental- vs. Kontrollgruppe), der Zeit (Vor- vs. Nachtest), der Klassenstufe (5. vs. 6. Klasse), der vier Antworten auf den beiden Aufgaben (Kühe Antwort 1 & 2, Spieleautomat Antwort 1 & 2) sowie eine Interaktion zwischen Bedingung und Zeit modelliert. Zusätzlich wurden random Intercepts der vier Kategoriengrenzen über Schulklassen hinweg modelliert (weitere random effects waren bei diesem Modellierungsansatz nicht implementierbar). In diesem Modell wurde mittels nominaler Effekte (auch kategorienspezifische Effekte genannt, siehe Bürkner and Vuorre 2019; Christensen 2018) zugelassen, dass sich Effekte der unabhängigen Variablen unterschiedlich auf die Kategoriengrenzen und damit die Wahrscheinlichkeiten auswirken, Argumentationen in den fünf Kategorien zu zeigen. Damit wurde der Annahme Rechnung getragen, dass sich das Training beispielsweise positiv auf die Wahrscheinlichkeit auswirken könnte, sich in der Argumentation auf Haupteffekte zu beziehen, während nicht damit gerechnet wurde, dass die Auswirkungen auf die Wahrscheinlichkeiten der anderen Kategorien ebenso positiv und genau gleich groß ausfallen würden. Lediglich für den Haupteffekt der Klassenstufe wurde ein genereller Effekt geschätzt, da hier keine Annahmen zu kategorienspezifischen Effekten vorlagen. Um p-Werte für die Haupt- und Interaktionseffekte zu ermitteln, wurden in einem schrittweisen Verfahren ausgehend vom beschriebenen Modell die einzelnen Effekte ausgeschlossen und die Passung dieser reduzierten Modelle mit dem vollen beschriebenen Modell anhand von chi^2-Unterschiedstests verglichen.

4 Ergebnisse

4.1 Evidenzbasierte Argumentation in 5. und 6. Klassenstufen im Vortest

Insgesamt wurden 2472 Antworten kodiert. Dies umfasste jeweils 1236 (zwei Antworten pro Aufgabe von $n = 618$ Schüler*innen) Antworten aus den Aufgaben „Kühe“ und „Spieleautomat“sowie 1920 Antworten (jeweils 4 Antworten von $n = 480$ Schüler*innen) aus der 5. Klassenstufe und 552 Antworten (jeweils 4 Antworten von $n = 138$ Schüler*innen) aus der 6. Klassenstufe. Fehlende Antworten, d. h. wenn die Schüler*innen im offenen Antwortformat nichts geschrieben hatten (aber im Single- Choice Teil eine Antwort gewählt hatten), wurden mit den Antworten der Kategorie *unverständliche Argumentation* zusammengelegt. Die Häufigkeit dieser Fälle war über die Aufgaben und Zeitpunkte hinweg vergleichbar gering ausgeprägt (VT Kühe: 1.9 %, VT Spieleautomat: 2.1 %, NT Kühe: 2.2 %, NT Spieleautomat: 2.7 %).

Abb. 4(a) zeigt die prozentuale Verteilung der Antworten in den fünf Evidenzkategorien im Vortest. Ebenso zeigt Abb. 4 die Verteilung aufgeteilt nach Aufgabe (B) und nach Klassenstufe (C). Zusätzlich sind in Tab. 5 (siehe Appendix) die entsprechenden Verteilungen aufgeteilt für die beiden Bedingungen ersichtlich.

Wie aus Abb. 4(a) hervorgeht, waren im Vortest etwa ein Viertel aller Schüler*innenantworten der Kategorie *unverständliche Argumentation* zuzuordnen; die niedrigste Kategorie war damit auch die am stärksten besetzte. Gleichzeitig wur-

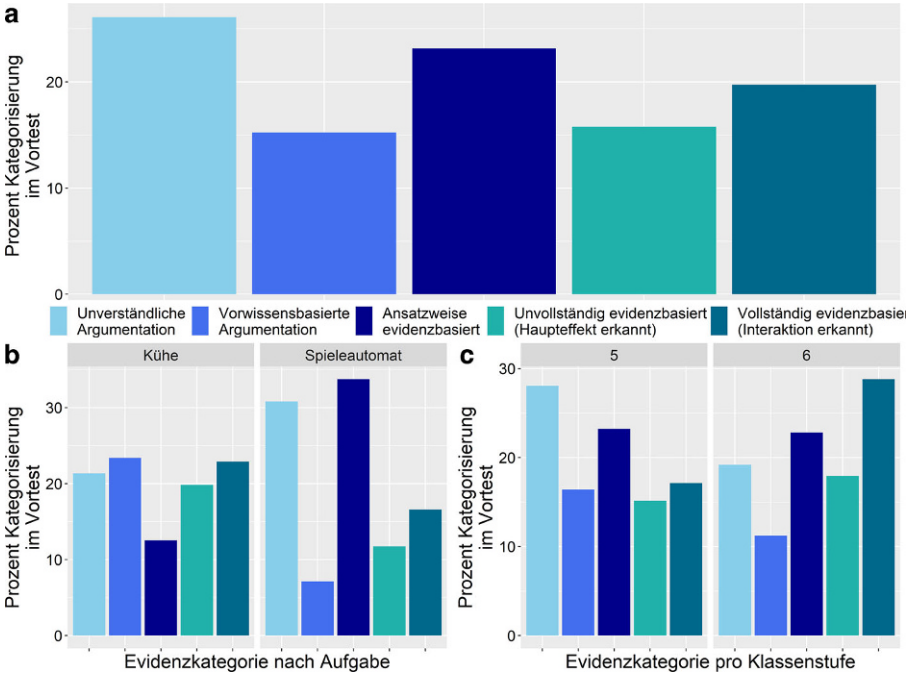


Abb. 4 Deskriptive Übersicht der prozentualen Verteilung der Argumentationen im Vortest

de jedoch in 20 % der Schüler*innenantworten die Interaktion bereits im Vortest korrekt erkannt und in der Argumentation beschrieben (*vollständig evidenzbasiert*). Insgesamt 59 % der Schüler*innenantworten waren zumindest ansatzweise evidenzbasiert.

Die Unterschiede zwischen den relativen Häufigkeiten der Antwortkategorien in Abb. 4(b) zeigen, dass sich die Argumentationen der Schüler*innen im Vortest je nach Aufgabe unterschieden. Insgesamt 62.0 % der Schüler*innen argumentierten bei der Spieleautomat-Aufgabe evidenzbasiert (berechnet aus Antworten, die mindestens der Kategorie *ansatzweise evidenzbasierte Argumentation* oder höher zuzuordnen sind) und 55.3 % bei der Kühe-Aufgabe. Ein chi^2-Test unterstreicht diesen Unterschied in der relativen Häufigkeit, mit der Schüler*innen bei den beiden Aufgaben evidenzbasiert argumentierten, als statistisch signifikant ($\chi^2 = 125.19$, df $= 1$, $p < 0.001$).

Zusätzlich zu diesem Unterschied in evidenzbasierter Argumentation zwischen den beiden Aufgaben wiesen die Schüler*innen bei der vorwissensnäheren Kühe-Aufgabe eine höhere prozentuale Ausprägung an vorwissensbasierten Argumentationen (23.4 %) auf als bei der vorwissensferneren Spieleautomat-Aufgabe (7.1 %). Die Kategorie der *vorwissensbasierten Argumentation* war bei der Kühe-Aufgabe die am stärksten besetzte und bei der Spieleautomat-Aufgabe die am niedrigsten besetzte. Fehlerhafte oder zu oberflächliche Antworten, welche der Kategorie der *ansatzweise evidenzbasierten Argumentation* zuzuordnen sind, waren bei der Spieleautomat-Aufgabe die am höchsten besetzte Kategorie (33.7 % aller Antworten) und in der Kühe-Aufgabe die am niedrigsten besetzte Kategorie (12.5 % aller Antworten). Während die Häufigkeit evidenzbasierter Argumentationen also bei der Spieleautomat-Aufgabe insgesamt höher ausgeprägt war als bei der Kühe-Aufgabe, waren die evidenzbasierten Argumentationen bei der Spieleautomat-Aufgabe häufiger falsch und somit mit inkorrekten Schlussfolgerungen bezüglich der Haupteffekte oder der Interaktion verknüpft.

Betrachtet man die beiden Klassenstufen getrennt (siehe Abb. 4c), so zeigten die Schüler*innen der 5. Klassenstufe insgesamt 55.5 % zumindest evidenzbasierte Argumentationen und die Schüler*innen der 6. Klassenstufe 69.6 %. Dieser Unterschied lässt sich mit einem chi^2-Test ($\chi^2 = 34.295$, df $= 1$, $p < 0.001$) als signifikant beschreiben. Dieses Ergebnis weist darauf hin, dass Schüler*innen der 6. Klassenstufe in stärkerem Ausmaß evidenzbasiert argumentieren als Schüler*innen der 5. Klassenstufe.

4.2 Effekte des VKS-Trainings auf die Qualität der evidenzbasierten Argumentation im „Interaktionseffekte"-Kontext

Um den Einfluss des VKS-Trainings auf die Qualität der evidenzbasierten Argumentation zu untersuchen, (Fragestellung 2) betrachteten wir zunächst deskriptiv, wie sich die Argumentationen anhand der einzelnen Kategorien des Kodierschemas vom Vortest zum Nachtest veränderten. Abb. 5 (siehe auch Tab. 5) zeigt die Veränderung der Häufigkeiten der einzelnen Evidenzkategorien über die Zeit, nach Experimental- und Kontrollgruppe aufgeteilt. In beiden Gruppen ist deskriptiv ein deutlicher Zuwachs der relativen Häufigkeit von Antworten in der Kategorie *unvoll-*

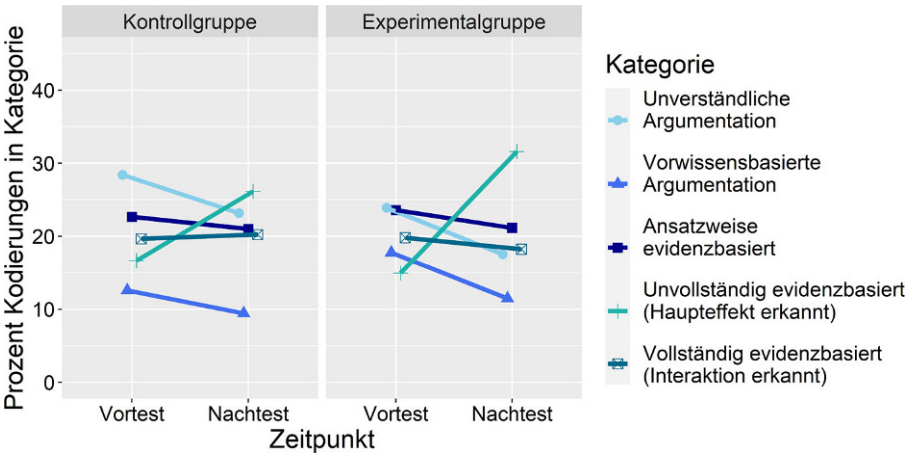


Abb. 5 Übersicht der Veränderung der einzelnen Kategorien von Vortest zu Nachtest, unterteilt in Unterschiede zwischen der Experimentalgruppe und der Kontrollgruppe sowie der Klassenstufe

ständig evidenzbasierte Argumentation (Haupteffekt erkannt) ersichtlich, welcher jedoch in der Experimentalgruppe noch einmal deutlich stärker ausgeprägt ist als in der Kontrollgruppe. Der Zuwachs in der relativen Häufigkeit dieser Kategorie geht in beiden Gruppen mit einer Abnahme der relativen Häufigkeiten fast aller anderen Kategorien einher. Auch die relative Häufigkeit der Argumentationskategorie *vollständig evidenzbasierte Argumentation* (Interaktion erkannt) geht in beiden Gruppen zurück. Ebenso geht der relative Anteil vorwissensbasierter Argumentationen in beiden Gruppen zurück, wobei dieser Rückgang in der Experimentalgruppe stärker scheint.

Um diese Eindrücke inferenzstatistisch abzusichern, wurde die ordinale Mehrebenen-Regressionsanalyse angepasst. Die Ergebnisse der Modellvergleichstests, welche die einzelnen Haupt- und Interaktionseffekte testen, sind in Tab. 3 berichtet. Sämtliche geschätzte Modellparameter sind in Tab. 4 berichtet. Die durch die geschätzten Modellparameter implizierten Wahrscheinlichkeiten der einzelnen Kategorien in den beiden Gruppen am Vor- und Nachtest entsprachen insgesamt weitest-

Tab. 3 Tests der Haupt- und Interaktionseffekte der ordinalen Mehrebenen-Regressionsanalyse zur Vorhersage der einzelnen Argumentationskategorien. Terme mit zwei Ausprägungen haben vier Freiheitsgrade, da außer des Effektes der Klassenstufe alle Effekte nominal (Kategorienspezifisch, d. h. ein Parameter für jeden der vier Kategorien-Thresholds) geschätzt wurden. Der Modellterm *Aufgabe* bildet Varianz in den Argumentationen der Schüler*innen über deren vier Antworten auf den beiden Aufgaben hinweg ab, somit hat dieser Term vier Ausprägungen und damit insgesamt zwölf Freiheitsgrade

Parameter	χ^2	df	p
Bedingung	28.55	4	< 0.001
Zeitpunkt	144.24	4	< 0.001
Bedingung*Zeitpunkt	11.78	4	0.019
Aufgabe	579.00	12	< 0.001
Klassenstufe	3.95	1	0.047

Tab. 4 Geschätzte Parameter, deren Standardfehler (SE) und resultierenden z- sowie *p*-Werte für die ordinale Mehrebenen-Regressionsanalyse. Die Intercepts beschreiben die Kategorien-Grenzen in der Vergleichsgruppe mit den Ausprägungen Antwort 1 der Kühe-Aufgabe, 5. Klassenstufe, Vortest, Kontrollgruppe Argumentationen in den fünf Kategorien zu zeigen. Für ein Beispiel der Verrechnung der geschätzten Modellparameter in implizierte vorhergesagte Wahrscheinlichkeiten, siehe Appendix. Für alle Effekte außer jenem der Klassenstufe, der nicht kategorienspezifisch geschätzt wurde, gibt es vier Parameter, welche kategorienspezifische Regressionsgewichte repräsentieren

	Parameter	SE	*z*	*p*
1|2 Intercept	−1.07	0.14	−7.44	< 0.001
2|3 Intercept	−0.28	0.14	−2.06	0.039
3|4 Intercept	0.32	0.14	2.33	0.020
4|5 Intercept	1.15	0.14	7.92	< 0.001
1|2 Kühe 2	−0.04	0.10	−0.38	0.701
2|3 Kühe 2	0.32	0.09	3.75	0.000
3|4 Kühe 2	0.27	0.08	3.19	0.001
4|5 Kühe 2	0.53	0.10	5.40	< 0.001
1|2 Spieleautomat 1	0.27	0.10	2.73	0.006
2|3 Spieleautomat 1	−0.21	0.09	−2.38	0.017
3|4 Spieleautomat 1	0.63	0.08	7.48	< 0.001
4|5 Spieleautomat 1	0.79	0.10	7.61	< 0.001
1|2 Spieleautomat 2	0.51	0.10	5.24	< 0.001
2|3 Spieleautomat 2	−0.11	0.09	−1.31	0.192
3|4 Spieleautomat 2	0.85	0.09	9.85	< 0.001
4|5 Spieleautomat 2	0.74	0.10	7.29	< 0.001
1|2 Zeitpunkt: Nachtest	−0.34	0.09	−3.77	0.000
2|3 Zeitpunkt: Nachtest	−0.38	0.09	−4.45	< 0.001
3|4 Zeitpunkt: Nachtest	−0.46	0.08	−5.42	< 0.001
4|5 Zeitpunkt: Nachtest	−0.03	0.10	−0.33	0.741
1|2 Bedingung: Experimentalgruppe	−0.23	0.09	−2.48	0.013
2|3 Bedingung: Experimentalgruppe	0.05	0.08	0.61	0.539
3|4 Bedingung: Experimentalgruppe	0.04	0.09	0.52	0.606
4|5 Bedingung: Experimentalgruppe	−0.02	0.10	−0.24	0.812
1|2 Interaktion: Nachtest × Experimentalgruppe	−0.11	0.13	−0.84	0.400
2|3 Interaktion: Nachtest × Experimentalgruppe	−0.25	0.12	−2.04	0.041
3|4 Interaktion: Nachtest × Experimentalgruppe	−0.20	0.12	−1.68	0.093
4|5 Interaktion: Nachtest × Experimentalgruppe	0.16	0.15	1.11	0.266
Klassenstufe 6	0.49	0.24	2.05	0.040

gehend den deskriptiv ersichtlichen Mustern, was eine gute Passung des Modells auf die Daten zeigte. Um die detaillierte Nachvollziehbarkeit des Modells und der Passung auf die Daten zu gewährleisten, sind im Appendix detailliertere deskriptive Statistiken über die Aufgaben und Klassenstufen hinweg, sowie ein Beispiel zur Interpretation der einzelnen Regressionsparameter gegeben.

Die Tests aus der ordinalen Regression in Tab. 3 bestätigen, dass Haupteffekte der Bedingung, des Zeitpunktes, der Aufgabe und der Klassenstufe, sowie eine Interaktion zwischen Bedingung und Zeitpunkt vorlagen. Der Haupteffekt der Aufgabe zeigt, dass die Wahrscheinlichkeiten, Argumentationen in den einzelnen Kategorien zu zeigen, zwischen den vier Antworten der Schüler*innen auf den beiden Aufgaben variierten. Der Haupteffekt der Klassenstufe bestätigt den schon für den Vortest beschriebenen Befund (siehe Abb. 4c), dass Schüler*innen der 6. Klassenstufe grundsätzlich in höherem Ausmaß evidenzbasiert argumentierten. Die positive Richtung dieses Effektes der Klassenstufe ist aus Tab. 4 ersichtlich.

Die Interaktion aus Bedingung und Zeit ist von zentralem Interesse für die Forschungsfrage und zeigt, dass sich die Kontroll- und Experimentalbedingung unterschiedlich auf die Veränderung in den Argumentationen der Schüler*innen vom Vor- zum Nachtest auswirkten. Die aus Abb. 5 ersichtlichen Effekte der Bedingung auf Veränderung in den Kategorien *vorwissensbasierte Argumentation* und *unvollständig evidenzbasierte Argumentation* werden aus den in Tab. 4 berichteten Ausprägungen der Interaktionseffekte aus Zeitpunkt und Bedingung (Nachtest × Experimentalgruppe) bestätigt. Das negative Regressionsgewicht für die Kategoriengrenze 2|3 zeigt, dass die Wahrscheinlichkeit einer Argumentation, die maximal in die Kategorie *vorwissensbasierte Argumentation* fällt, in der Experimentalgruppe stärker abnahm als in der Kontrollgruppe. Während das negative Regressionsgewicht für die Interaktion von −0.20 auf die Kategoriengrenze 3|4 zeigt, dass Argumentationen, welche höchstens der Kategorie *ansatzweise evidenzbasierte Argumentation* zuzuschreiben sind, in der Experimentalgruppe stärker abnahmen als in der Kontrollgruppe, zeigt das positive Regressionsgewicht für die Kategoriengrenze 4|5, dass gleichzeitig die Anzahl der in der vierten Kategorie verordneten Argumentationen in der Experimentalgruppe stärker zunahm als in der Kontrollgruppe.

5 Diskussion

Ziel dieser Studie war die Untersuchung des Grades und der Qualität der Evidenzbasierung bei der Argumentation von Schüler*innen der 5. und 6. Klassenstufe, die gezogene Schlussfolgerungen aus Aufgaben mit einem „Interaktionseffekte"-Kontext anhand vorgegebener, experimenteller Daten begründen sollten. Des Weiteren war von Interesse, ob ein explizites Training der Variablenkontrollstrategie, dessen Wirksamkeit auf evidenzbasierte Argumentation in einem „Haupteffekte"-Kontext bereits gezeigt werden konnte (Peteranderl 2019), sich auch auf die evidenzbasierte Argumentation der Schüler*innen in einem „Interaktionseffekte"-Kontext auswirkt. Zusätzlich wurden als weitere potenzielle Einflussfaktoren das Alter der Schüler*innen (Klassenstufe) sowie der Aufgabenkontext untersucht.

5.1 Fähigkeit von 5./6.-Klässler*innen zur evidenzbasierten Argumentation in einem „Interaktionseffekte"- Kontext

Unsere Ergebnisse zeigen, dass Kinder der Klassenstufe 5 und 6 durchaus bereits zu evidenzbasierten Argumentationen in einem „Interaktionseffekte"-Kontext in der

Lage sind. Auch wenn viele Antworten noch unverständliche oder fehlerhafte Argumentationsversuche enthielten, beurteilten doch bereits etwa 20 Prozent der Schüler*innen zum Vortestzeitpunkt den zu interpretierenden Interaktionseffekt korrekt. Bei älteren Schüler*innen (6. Klassenstufe) wurde im Vergleich zu jüngeren Schüler*innen (5. Klassenstufe) eine höhere Häufigkeit fortgeschrittener Argumentationen festgestellt: Die jüngeren Schüler*innen argumentierten unsystematischer und vermehrt vorwissensbasiert. Die älteren Schüler*innen zeigten weniger vorwissensbasierte Argumentationen und einen höheren Grad an evidenzbasierter Argumentation. Dieses Ergebnis passt zu Kuhns Modell der Entwicklung des Experimentierverständnisses (Kuhn 2000, 2002). Nach diesem Modell nehmen jüngere Kinder Theorie und Evidenz noch nicht als zwei getrennte Phänomene wahr und entwickeln die notwendige Abgrenzung und Koordination von Theorie und Evidenz erst mit zunehmendem Alter. Ein weiterer Grund könnte sein, dass ältere Schüler*innen ein allgemein besser ausgebildetes Ausdrucksvermögen oder eine allgemein bessere Argumentationskompetenz besitzen.

In Bezug auf unsere erste Fragestellung lässt sich zunächst einmal festhalten, dass 5.- und 6.-Klässler*innen Interaktionseffekte prinzipiell bereits erkennen und ihre Schlussfolgerungen evidenzbasiert begründen können. Dies gelingt älteren Schüler*innen im Schnitt besser als jüngeren. Viele prinzipiell evidenzbasierte Argumentationen von Schüler*innen dieser Altersgruppe fokussieren allerdings nur einen der Haupteffekte und lassen die qualifizierende Interaktion außer Acht. Wie auch im klassischerweise untersuchten „Haupteffekte"-Kontext, greifen Schüler*innen, die nicht evidenzbasiert argumentieren, oft auf ihre persönlichen Theorien und Überzeugungen zurück.

5.2 Trainingseffekte

Diese Studie evaluierte die Effekte eines umfangreichen, alle vier Facetten der VKS umfassenden Trainings, welches jedoch in den klassischerweise zur Einführung der VKS-Strategie verwendeten „Haupteffekte"-Kontext eingebettet war. Das Training verbesserte zwar die Fähigkeit der Schüler*innen, sich in ihrer Argumentation überhaupt auf Evidenz zu beziehen. Allerdings bereitete es sie nicht darauf vor, Daten im Sinn eines Interaktionseffektes zu interpretieren und dementsprechend zu argumentieren. Auch wenn der deutlichste Zuwachs an evidenzbasierter Argumentation in der Experimentalgruppe (mit VKS-Training) erfolgte, so war doch auch in der Kontrollgruppe (mit rein inhaltlich ausgerichtetem Stromkreise-Training) eine Erhöhung des Ausmaßes an evidenzbasierter Argumentation vom Vor- zum Nachtest zu beobachten. Da weder die VKS noch allgemein evidenzbasierte Argumentation Bestandteile des Kontrolltrainings waren, ist ein Trainingseffekt hier unwahrscheinlich. Auch erhielten die Lehrpersonen der Kinder zwar Unterrichtsmaterialien für beide Trainings, um diese, falls gewünscht, nachträglich allen Kindern ihrer Klasse zu vermitteln; sie wurden jedoch gebeten, damit in jedem Fall bis zur Beendigung des Nachtests zu warten. Es kann nicht ganz ausgeschlossen werden, dass die Kinder einer Klasse sich untereinander über die jeweiligen Trainingsinhalte austauschten; dies geschah jedoch vermutlich nicht in systematischer und strukturierter Form. Bei der wiederholten Erfassung von Fähigkeiten, welche auf der Variablenkontrolle basieren,

zeigten sich auch in anderen Stichproben in der Vergangenheit Zunahmen (Bohrmann 2017; Schalk et al. 2019; Tempel et al. 2020). Neben allgemeinen Effekten durch kognitive Entwicklungsprozesse, die wir für unsere Studie angesichts der relativ kurzen Zeitspanne ausschließen möchten, handelt es sich dabei in der Regel um Retest-Effekte (Ferrer et al. 2004; Tempel et al. 2020). Bei der zweiten Bearbeitung des Tests könnten die Schüler*innen beispielsweise durch Wiedererkennen der Aufgabeninhalte weniger Aufwand für das inhaltliche Verständnis der Aufgaben benötigt haben, wodurch sie sich mehr auf die Eigenschaften der dargestellten Vergleiche konzentrieren konnten. Dies könnte dazu geführt haben, dass in der Kontrollgruppe die Häufigkeiten unverständlicher sowie vorwissensbasierter Argumentationen abnahmen und die Häufigkeit der Argumentationen, in welchen Haupteffekte erkannt wurden, zunahm. Die Zunahme evidenzbasierter Argumentationen in der Kontrollgruppe war jedoch weniger stark als in der Interventionsgruppe. Dies spricht für einen klaren, über den Retest-Effekt hinausgehenden Effekt des VKS-Trainings.

Das VKS-Training führte zwar zu einer Reduzierung des Anteils vorwissensbasierter Begründungen und einer Erhöhung evidenzbasierter Begründungen. Jedoch zeigte sich keine Erhöhung vollständig evidenzbasierter Begründungen (Begründungen, welche die Interaktion in den Daten berücksichtigen). Der Grund dafür ist vermutlich die Ausrichtung des Trainings auf den klassischerweise zur Testung und Vermittlung der VKS eingesetzten „Haupteffekte"-Kontext. Die Schüler*innen lernten während des Trainings zwar, Vergleiche zwischen Experimentdurchgängen zu ziehen und jeweils einzelne Variablen separat zu überprüfen; sie lernten aber nicht, für alle Variablen systematisch auch Zusammenhänge untereinander zu testen, bevor sie die Ergebnisse interpretieren. Während dies im „Haupteffekte"-Kontext ein zielführendes und sparsames Vorgehen ist, werden durch die fehlende Überprüfung von Effekten eines Faktors auf unterschiedlichen Stufen weiterer Faktoren bestehende Interaktionen übersehen.

Ein weiterer Grund, weshalb in der vorliegenden Studie kein Transfer auf den „Interaktionseffekte"- Kontext gefunden werden konnte, könnte sein, dass Schüler*innen bei herausfordernden Aufgaben größere Schwierigkeiten haben, die Qualität ihrer Argumentation zu verbessern (vlg. McNeill 2011). Laut Greiff et al. (2012) erhöht sich die Schwierigkeit einer Aufgabe mit der Komplexität des Kausalmodells. Eine systematische Interaktion zwischen zwei Variablen erhöht die Komplexität des Kausalmodells der Variablen in einer Aufgabe und damit auch die Schwierigkeit der Untersuchung der einzelnen Variablen anhand der VKS.

In Bezug auf unsere zweite Fragestellung lässt sich demnach festhalten, dass das im „Haupteffekte"- Kontext angesiedelte VKS-Training zwar den Anteil evidenzbasierter Argumentationen in den Antworten der Kinder insgesamt substantiell erhöhte, nicht jedoch die spezifische Fähigkeit, Interaktionseffekte argumentativ zu berücksichtigen. Angesichts der Tatsache, dass etliche Schüler*innen diese Fähigkeit jedoch bereits ohne spezifisches Training zeigten, wäre zu überlegen, ob ein klassisches VKS-Training wie das hier beschriebene mit einer Ergänzung um einige Aufgaben aus einem „Interaktionseffekte"- Kontext nicht bereits deutliche positive Effekte auch auf die Fähigkeit der Kinder zur evidenzbasierten Argumentation in Aufgaben wie den hier vorgestellten zeigen könnte.

5.3 Aufgabeneffekte

Unsere Analysen zeigten eine Abhängigkeit der von den Kindern eingesetzten Argumentationsstrategien von der Einbettung der Aufgabe, wobei allerdings mehrere potentiell relevante Aufgabendimensionen (Vorwissensnähe und Art der Interaktion) konfundiert waren und unsere Schlussfolgerungen hier daher nur tentativ sein können. Wir analysierten die Antworten von Kindern im Rahmen von zwei Aufgabenkontexten: einer vorwissensnäheren Aufgabe mit ordinaler Interaktion („Kühe“) und einer abstrakteren und dadurch vorwissensferneren Aufgabe mit disordinaler Interaktion („Spieleautomat“). Unsere Ergebnisse zeigten, dass die Schüler*innen der 5. und 6. Klassenstufe bei der Kühe-Aufgabe häufiger auf ihr konzeptuelles Vorwissen zurückgriffen, als sie es bei der vorwissensferneren Spieleautomat-Aufgabe taten. Wenn sie jedoch bei der Kühe-Aufgabe evidenzbasiert argumentierten (also ihre Schlussfolgerungen anhand der Daten begründeten), dann erkannten sie mit größerer Wahrscheinlichkeit auch die Interaktion. Diese Effekte erlauben aufgrund der Konfundierung von Vorwissensnähe und Interaktionstyp keine eindeutigen Schlussfolgerungen. Jedoch unterstützen unsere Ergebnisse in Verbindung mit der bisherigen Forschung (Croker and Buchanan 2011; Kuhn et al. 1988; Lazonder et al. 2008; Wilhelm and Beishuizen 2003) die Annahme, dass sich Vorwissen auch im „Interaktionseffekte“-Kontext auf die Qualität von evidenzbasierten Begründungen auswirkt. Um die Effekte des Aufgabenkontextes detaillierter zu untersuchen, sollten in zukünftigen Untersuchungen potentiell relevante Aufgabendimensionen systematisch variiert werden: Beispielsweise die Vorwissensnähe (des in der Aufgabe experimentell untersuchten Sachgebiets), die Art der Interaktion (ordinal, disordinal) und die Komplexität (z. B. über die Anzahl der zu berücksichtigenden unabhängigen Variablen). Eine Möglichkeit für ein Forschungsprogramm wäre beispielsweise, zunächst nur Effekte der zu untersuchenden Interaktion (ordinal vs. disordinal) anhand jeweils mehrerer Aufgaben desselben Interaktionstyps zu untersuchen. In weiteren Schritten könnten dann weitere Einflussfaktoren (Vorwissensnähe, Aufgabenkomplexität, etc.) einbezogen und der Fragenkatalog dementsprechend ergänzt werden.

5.4 Fazit

Bisherige Forschung zur Erfassung und Förderung der VKS nutzte ausschließlich „Haupteffekte“-Kontexte (Keselman 2003; Kuhn et al. 2000, 2008, 2009). Unsere Studie zeigt, dass ein Transfer auf den „Interaktionseffekte“-Kontext nicht ohne Weiteres erwartet werden kann. Zukünftige Studien sollten im Rahmen eines größeren Aufgabenumfanges einerseits systematisch untersuchen, welche Rolle Aspekte wie Komplexität (z. B. Anzahl der Variablen), Vorwissensnähe und Art der Interaktion für das Erkennen von Interaktionseffekten spielen. Andererseits sollte untersucht werden, welche Arten von Lernumgebungen dazu beitragen können, dass Interaktionen zwischen Variablen korrekt erkannt und interpretiert werden können. Daraus könnten für den naturwissenschaftlichen Unterricht praxisnahe und lohnenswerte Implikationen abgeleitet werden. Eine Möglichkeit wäre, die bereits bestehende theoretische und experimentelle Forschung zum Erwerb der Fähigkeiten und Kompetenzen für wissenschaftliches Denken und experimentelle Hypothesentestung auf die Unter-

suchung von Kausalzusammenhängen, die über additive, konsistente Haupteffekte hinausgehen, im Schulkontext zu erweitern. Dabei würde man sich gegebenenfalls von der expliziten Instruktion der VKS als zentralem Lerninhalt entfernen (vgl. Kuhn et al. 2008) und diese durch Bausteine ergänzen, die Kausalmodelle und Begründungen im „Interaktionseffekte"-Kontext explizit thematisieren.

6 Danksagung

Dieses Forschungsprojekt wurde durch einen ETH Research Grant und die ETH Foundation (Grant No. ETH-23 15-1) unterstützt. Die Autoren bedanken sich für die Unterstützung durch Elsbeth Stern und Ralph Schumacher (ETH Zürich), die die Durchführung im Rahmen der Schweizer MINT-Studie ermöglichten und unterstützten. Außerdem bedanken sich die Autoren bei allen Lehrpersonen und Klassen, die an der Studie teilgenommen haben sowie bei Yvonne Oberholzer für ihren engagierten Einsatz bei der Entwicklung des Kodiermanuals.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich

Appendix

Kategorienhäufigkeiten im Vor- und Nachtest nach Bedingung, Klassenstufe und Aufgabe: Interpretationsbeispiel für die ordinale Regressionsanalyse

Im Folgenden wird anhand eines Beispiels erläutert, wie die in Tab. 4 ersichtlichen Parameter aus der ordinalen Regressionsanalyse interpretiert werden. Dabei wird den Leser*innen anhand eines konkreten Rechenbeispiels erläutert, wie die Wahrscheinlichkeiten für jede Antwort berechnet werden, eine spezifische Ausprägung zu zeigen (in eine spezifische Kategorie des Kodierschemas – abhängig von den Wahrscheinlichkeiten, in die jeweils anderen Kategorien zu fallen). Dazu ist wichtig zu wissen, dass die in Tab. 4 ersichtlichen Intercepts auf den Variablenausprägungen Vortest, Kontrollgruppe und Antwort 1 auf der Kühe-Aufgabe beruhen. Diese Bedingungen stellen, ähnlich wie in einer Varianzanalyse, die Vergleichsgruppe dar, anhand de-

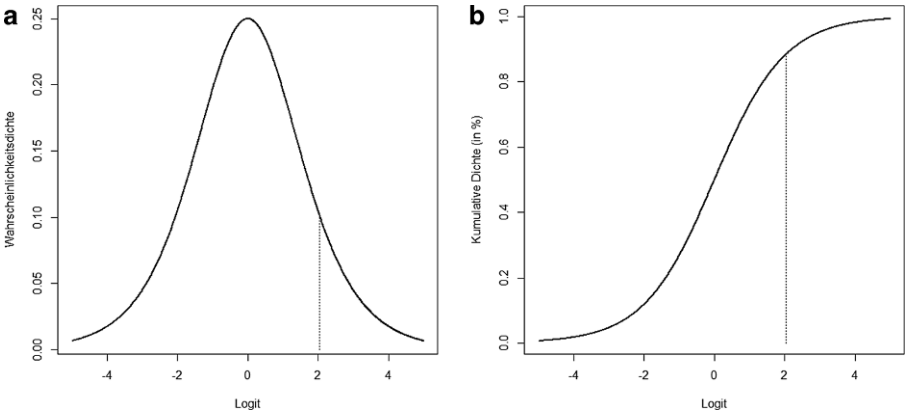


Abb. 6 Darstellung der Dichte der standard-logistischen Verteilung **(a)** auf welcher die Kategoriengrenzen der ordinalen Regression aufgetragen werden sowie der resultierenden kumulativen Dichte **(b)**. In *gestrichelter Linie* ist exemplarisch eine Kategoriengrenze mit einem Logit von 2.04 aufgetragen

rer die anderen Parameter im Modell skaliert und deren Signifikanzen berechnet werden. Weiterhin ist relevant zu wissen, dass alle Modellparameter sogenannten Logits (log-odds) entsprechen, die mittels $\exp(\text{parameter})/(1 + \exp(\text{parameter}))$ in Wahrscheinlichkeiten umgerechnet werden können. Um Leser*innen nun bei der Interpretation und der Nachvollziehbarkeit der Ergebnisse zu unterstützen, wird hier illustriert, wie diese Umrechnung in Wahrscheinlichkeiten bei bestimmten Ausprägungen der Modellparameter durchgeführt wird. Beispielsweise kann die Modell-implizierte Wahrscheinlichkeit einer Argumentation, welche in die Kategorie *unvollständig evidenzbasiert (Haupteffekt erkannt)* in der Experimentalgruppe am Nachtest bei der Spieleautomat-Aufgabe fällt, folgendermaßen berechnet werden:

Hierfür werden die Wahrscheinlichkeiten auf den beiden Antworten auf dieser Aufgabe einzeln berechnet und dann gemittelt. Die entsprechende Wahrscheinlichkeit für die erste Antwort auf der Spieleautomat-Aufgabe wird erhalten, indem man zum Intercept für die Kategoriengrenze 4|5 (1.15; dieser Parameter entspricht den Log-Odds der zweithöchsten im Vergleich zur höchsten Kategorie), welcher für die Kontrollgruppe am ersten Messzeitpunkt auf der ersten Antwort auf der Kühe-Aufgabe steht, erst das Logit-Regressionsgewicht für die entsprechende Kategoriengrenze der ersten Antwort auf der Spieleautomat-Aufgabe (0.79) addiert. Danach werden noch die Logit-Regressionsgewichte für den Nachtest (−0.03), die Experimentalgruppe (−0.02), sowie die Interaktion aus Zeit und Bedingung (0.16) addiert. Durch diese Aufaddierung, die gleich wie in einem regulären multiplen Regressionsmodell erfolgt, kommt man an den resultierenden geschätzten Logit-Parameter für die gewünschten Ausprägungen der Prädiktorvariable. Der resultierende Logit unter diesen Bedingungen von $1.15 + 0.79 - 0.03 - 0.02 + 0.16 = 2.04$ wird per $\exp(2.04)/(1 + \exp(2.04))$ in eine Wahrscheinlichkeit von 0.88 umgerechnet. Diese Wahrscheinlichkeit sagt aus, dass Schüler*innen mit einer Wahrscheinlichkeit von 88 % bei dieser Antwort maximal die Kategorie 4 zeigen (das wäre dann die Interpretation). Diese Umrechnung ist in Abb. 6 dargestellt. In Abb. 6a ist der Logit von 2.04

Tab. 5 Wahrscheinlichkeiten in Prozenten, in die jeweiligen Kategorien des Kodierschemas zu fallen; abhängig von Zeit, Bedingung, Klassenstufe und Aufgabe

Zeit	Bedingung	Klassenstufe	Aufgabe	% Unverständliche Argumentation	% Vorwissensbasierte Argumentation	% Ansatzweise evidenzbasiert	% Unvollständig evidenzbasiert (Haupteffekt erkannt)	% Vollständig evidenzbasiert (Interaktion erkannt)
Vortest	KG	5	Kühe	24	23	13	19	20
Vortest	KG	5	Spieleautomat	36	5	31	13	15
Nachtest	KG	5	Kühe	23	16	12	26	23
Nachtest	KG	5	Spieleautomat	26	5	32	23	14
Vortest	EG	5	Kühe	22	28	12	20	18
Vortest	EG	5	Spieleautomat	30	10	37	9	15
Nachtest	EG	5	Kühe	17	21	13	29	21
Nachtest	EG	5	Spieleautomat	20	6	33	29	12
Vortest	KG	6	Kühe	18	14	14	20	34
Vortest	KG	6	Spieleautomat	25	3	34	19	19
Nachtest	KG	6	Kühe	13	11	10	34	32
Nachtest	KG	6	Spieleautomat	25	2	26	28	18
Vortest	EG	6	Kühe	12	18	12	21	38
Vortest	EG	6	Spieleautomat	22	10	32	12	24
Nachtest	EG	6	Kühe	12	7	10	40	31
Nachtest	EG	6	Spieleautomat	16	3	22	42	17

KG = Kontrollgruppe, EG = Experimentalgruppe

auf der Dichtefunktion der standard-logistischen Variable aufgetragen. Der Anteil der Dichte, welcher unter (links von) dem Logit von 2.04 liegt, entspricht folglich der Wahrscheinlichkeit, eine Antwort zu zeigen, die maximal in die Kategorie 4 fällt. In Abb. 6b ist aufgetragen, dass dies 88 % (ersichtlich anhand der kumulativen Dichte, welche auf der y-Achse abgetragen ist) der Fläche der Verteilung (kumulativen Dichte der Verteilung) entspricht. Ebenso wird anschließend die Wahrscheinlichkeit, maximal die Kategorie 3 zu zeigen, anhand derselben Parameter, nun allerdings immer eine Kategoriengrenze tiefer (3|4 anstatt 4|5) berechnet, was zu einem Logit von $0.32 + 0.63 - 0.46 + 0.04 - 0.20 = 0.68$ führt, der in eine Wahrscheinlichkeit von 66 % umgerechnet wird. Diese beiden Wahrscheinlichkeiten werden nun voneinander abgezogen, um die Wahrscheinlichkeit zu erhalten, eine Argumentation zu zeigen, die genau in die höhere der beiden Kategorien fällt. $88\,\% - 66\,\%$ ergeben eine Wahrscheinlichkeit von 22 %, unter den erwähnten Bedingungen Kategorie 4 zu zeigen. Dasselbe wird dann für Antwort 2 auf der Spieleautomat-Aufgabe gemacht. Für Antwort 2 ergibt sich eine Wahrscheinlichkeit von 27 %. Der Mittelwert dieser beiden errechneten Wahrscheinlichkeiten beträgt $(22\,\% + 27\,\%)/2 = 24.5\,\%$, was in etwa der beobachteten Wahrscheinlichkeit, auf der Spieleautomat-Aufgabe am Nachtest in der Experimentalgruppe eine Argumentation zu zeigen, die in die Kategorie *unvollständig evidenzbasiert (Haupteffekt erkannt)* fällt, entspricht (siehe Tab. 5).

Literatur

Bohrmann, M. (2017). *Zur Förderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht*. Berlin: Logos.

Budke, A., & Meyer, M. (2015). *Fachlich argumentieren lernen – Die Bedeutung der Argumentation in den unterschiedlichen Schulfächern*. Münster: Waxmann.

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.

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Hrsg.), *Human development from early childhood to early adulthood: findings from a 20 year longitudinal study* (S. 173–197). Hove: Psychology Press.

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: a tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199.

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.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction. *Review of Educational Research*, *63*(1), 1–49. https://doi.org/10.3102/00346543063001001.

Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: how is conceptual change impeded? *Journal of Educational Psychology*, *94*(2), 327–343. https://doi.org/10.1037/0022-0663.94.2.327.

Christensen, R. H. B. (2015). *Analysis of ordinal data with cumulative link models—estimation with the r-package ordinal. R-package version, 28*

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the r package ordinal. Submitted in *J. Stat. Software*.

Croker, 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*(3), 409–424. https://doi.org/10.1348/026151010X496906.

Dewey, J. (2002). *Logik: die Theorie der Forschung*. Berlin: Suhrkamp.
Driver, R., Newton, P., & Osborne, J. F. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A.
Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046.
D-EDK (2016). Lehrplan 21 – Natur, Mensch, Gesellschaft. Bereinigte Fassung vom 29.02.2016. Deutschschweizer Erziehungsdirektoren-Konferenz, Luzern. http://v-ef.lehrplan.ch/downloads.php. Zugegriffen: 16. Okt. 2020
Edelsbrunner, P. A. (2017). *Domain-general and domain-specific scientific thinking in childhood: measurement and educational interplay*. PhD thesis, ETH Zürich.
Feißt, M., Hennigs, A., Heil, J., Moosbrugger, H., Kelava, A., Stolpner, I., Kieser, M., & Rauch, G. (2019). Refining scores based on patient reported outcomes–statistical and medical perspectives. *BMC Medical Research Methodology*, *19*, 167. https://doi.org/10.1186/s12874-019-0806-9.
Ferrer, E., Salthouse, T. A., Stewart, W. F., & Schwartz, B. S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging*, *19*(2), 243. https://doi.org/10.1037/0882-7974.19.2.243.
Gopnik, A., & Schulz, L. E. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176. https://doi.org/10.1037/0012-1649.40.2.162.
Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Applied Psychological Measurement*, *36*(3), 189–213. https://doi.org/10.1177/0146621612439620.
Gwet, K. L. (2014). *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. Gaithersburg: Advanced Analytics.
Haslbeck, H. (2019). *Die Variablenkontrollstrategie in der Grundschule*. PhD thesis, Technische Universität München.
Honda, C., & Ohyama, T. (2020). Homogeneity score test of ac 1 statistics and estimation of common ac 1 in multiple or stratified inter-rater agreement studies. *BMC medical research methodology*, *20*, 20. https://doi.org/10.1186/s12874-019-0887-5.
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: from childhood to adolescence*. New York: Basic Books.
Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: an analysis of university oceanography students' use of evidence in writing. *Science Education*, *86*(3), 314–342. https://doi.org/10.1002/sce.10024.
Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, *40*(9), 898–921. https://doi.org/10.1002/sce.10024.
Kim, H., & Song, J. (2006). The features of peer argumentation in middle school students' scientific inquiry. *Research in Science Education*, *36*(3), 211–233. https://doi.org/10.1007/s11165-005-9005-2.
Kind, P. M., Kind, V., Hofstein, A., & Wilson, J. (2011). Peer argumentation in the school science laboratory—exploring effects of task features. *International Journal of Science Education*, *33*(18), 2527–2558. https://doi.org/10.1080/09500693.2010.550952.
Klahr, D. (2000). *Exploring science: the cognition and development of discovery processes*. Cambridge: MIT Press.
Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: a developmental study. *Cognitive Psychology*, *25*(1), 111–146. https://doi.org/10.1006/cogp.1993.1003.
Koslowski, B. (2012). Scientific reasoning: explanation, confirmation bias, and scientific practice. In G. Feist & M. E. Gorman (Hrsg.), *Handbook of the psychology of science* (S. 151–192). Heidelberg: Springer.
Kruglanski, A., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(2), 97–109. https://doi.org/10.1037/a0020762.
Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science*, *9*(5), 178–181. https://doi.org/10.1111/1467-8721.00088.
Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Hrsg.), *Blackwell handbook of childhood cognitive development* (S. 371–393). New Jersey: Wiley-Blackwell. https://doi.org/10.1111/j.1467-9280.2005.01628.x.
Kuhn, D. (2007). Reasoning about multiple variables: control of variables is not the only challenge. *Science Education*, *91*(5), 710–726. https://doi.org/10.1002/sce.20214.

Kuhn, D. (2011). What is scientific thinking and how does it develop? In U. Goswami (Hrsg.), *Blackwell Handbook of Childhood Cognitive Development* (S. 497–523). New-Jersey: Wiley-Blackwell.

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. Cambridge: Academic Press.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*(4), 495–523. https://doi.org/10.1207/S1532690XCI1804_3.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S.H., Klahr, D., & Carver, S.M. (1995). *Strategies of knowledge acquisition*. Monographs of the Society for Research in Child Development. (S. i–157). https://doi.org/10.2307/1166059.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451. https://doi.org/10.1016/j.cogdev.2008.09.006. scientific reasoning – Where are we now?.

Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating the effects of multiple variables: a skill fundamental to scientific thinking. *Journal of Experimental Child Psychology*, *103*(3), 268–284. https://doi.org/10.1016/j.jecp.2009.01.009.

Lazonder, A.W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366.

Lazonder, A.W., Wilhelm, P., & Hagemans, M.G. (2008). The influence of domain knowledge on strategy use during simulation-based inquiry learning. *Learning and Instruction*, *18*(6), 580–592. https://doi.org/10.1016/j.learninstruc.2007.12.001.

Ludwig, T. (2017). *Argumentieren beim Experimentieren in der Physik – Die Bedeutung personaler und situationaler Faktoren*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät. https://doi.org/10.18452/18408.

Lunetta, V.N., Hofstein, A., & Clough, M.P. (2007). Learning and teaching in the school science laboratory: an analysis of research, theory, and practice. *Handbook of research on science education*, *2*, 393–441.

Masnick, A.M., & Klahr, D. (2003). Error matters: an initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development*, *4*(1), 67–98. https://doi.org/10.1080/15248372.2003.9669683.

Masnick, A., Klahr, D., & Knowles, E. (2002). Data-driven belief revision in children and adults. *Journal of Cognition and Development*, *18*(1), 87–109.

McNeill, K.L. (2011). Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *Journal of Research in Science Teaching*, *48*(7), 793–823. https://doi.org/10.1002/tea.20430.

National Research Council (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, D.C.: National Academies Press.

Osborne, J.F. (2010). Arguing to learn in science: the role of collaborative, critical discourse. *Science*, *328*(5977), 463–466.

Osborne, J.F., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020. https://doi.org/10.1002/tea.20035.

Osborne, J.F., Henderson, B., MacPherson, A., Szu, E., Wild, A., & Yao, S.Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846. https://doi.org/10.1002/tea.21316.

Osterhaus, C., Koerber, S., & Sodian, B. (2015). Children's understanding of experimental contrast and experimental control: an inventory for primary school. *Frontline Learning Research*, *3*(4), 56–94.

Peteranderl, S. (2019). *Experimentation skills of primary school children*. PhD thesis, ETH Zurich.

Revelle, W. (2013). *Using R and the psych package to find ω*

Rod Watson, J., Swain, J.R., & McRobbie, C. (2004). Students' discussions in practical scientific inquiries. *International Journal of Science Education*, *26*(1), 25–45. https://doi.org/10.1080/0950069032000072764.

Rosseel, Y. (2012). Lavaan: an r package for structural equation modeling and more. Version 0.5–12 (beta). *Journal of Statistical Software*, *48*(2), 1–36.

Rottman, B.M., Gentner, D., & Goldwater, M.B. (2012). Causal systems categories: differences in novice and expert categorization of causal phenomena. *Cognitive Science*, *36*(5), 919–932. https://doi.org/10.1111/j.1551-6709.2012.01253.x.

Ryu, S., & Sandoval, W.A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, *96*(3), 488–526. https://doi.org/10.1002/sce.21006.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472.

Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learning and Instruction*, *59*, 34–45. https://doi.org/10.1016/j.learninstruc.2018.09.006.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119. https://doi.org/10.1037/0012-1649.32.1.102.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*(9), 859–882. https://doi.org/10.1002/tea.3660280910.

Schwichow, M., Christoph, S., Boone, W., & Härtig, H. (2016). 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., 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*, 923. https://doi.org/10.1016/j.cedpsych.2020.101923.

Siegler, R. S., Liebert, D. E., & Liebert, R. M. (1973). Inhelder and piaget's pendulum problem: teaching preadolescents to act as scientists. *Developmental Psychology*, *9*(1), 97.

Siler, S. A., & Klahr, D. (2012). Detecting, classifying, and remediating: children's explicit and implicit misconceptions about experimental design. In R. W. P. E. J. Capaldi (Hrsg.), *Psychology of science: Implicit and explicit processess* (S. 137). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199753628.003.0007.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753–766. https://doi.org/10.1111/j.1467-8624.1991.tb01567.x.

Studhalter, U. T., Leuchter, M., Tettenborn, A., Elmer, A., Edelsbrunner, P. A., & Saalbach, H. (2021). Early science learning: the effects of teacher talk. *Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2020.101371.

Tempel, T., Kaufmann, K., Kranz, J., & Möller, A. (2020). Retrieval-based skill learning: testing promotes the acquisition of scientific experimentation skills. *Psychological Research*, *84*(3), 660–666. https://doi.org/10.1007/s00426-018-1088-2.

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Development*, *51*, 1–10.

Wilhelm, P., & Beishuizen, J. (2003). Content effects in self-directed inductive learning. *Learning and Instruction*, *13*(4), 381–402. https://doi.org/10.1016/S0959-4752(02)00013-0.

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

Zimmerman, C., & Croker, S. (2013). *Learning science through inquiry*. New York: Springer.

Zimmerman, C., & Klahr, D. (2018). *Development of scientific thinking. Stevens' handbook of experimental psychology and cognitive neuroscience*. Bd. 4 (S. 1–25). https://doi.org/10.1002/9781119170174.epcn407.