



– Originalbeitrag –

---

# Entwicklung und Validierung eines Messinstruments zur Erfassung systemischen Denkens zum Klimawandel bei Grundschülerinnen und Grundschülern

Sven Frey<sup>1</sup>, Maik Beege<sup>2</sup>, Nadine Tramowsky<sup>1</sup> und Werner Rieß<sup>1</sup>

<sup>1</sup>Pädagogische Hochschule Freiburg,  
Institut für Biologie und ihre Didaktik

<sup>2</sup>Pädagogische Hochschule Freiburg,  
Institut für Psychologie

---

## ZUSAMMENFASSUNG

Im Kontext der globalen Herausforderung des Klimawandels sollen Schülerinnen und Schüler im Rahmen einer Bildung für nachhaltige Entwicklung (BNE) dazu befähigt werden, komplexe Zusammenhänge zu verstehen. Dafür ist es sinnvoll, ihr systemisches Denken zu fördern, also die Fähigkeit, komplexe Wirklichkeitsbereiche als Systeme zu erkennen, zu beschreiben und zu modellieren. Um die Wirksamkeit von Unterrichtsmaßnahmen zur Förderung systemischen Denkens überprüfen zu können, bedarf es eines passenden und handhabbaren Diagnoseinstruments. Ein solches Instrument, welches das Lösen komplexer Probleme mit Systemmodellen zum Klimawandel bereits bei Grundschülerinnen und Grundschülern erfasst, gibt es bislang nicht. Im Rahmen des Projektes SysDeKlima wurde deshalb auf Grundlage eines in höheren Jahrgangsstufen bewährten heuristischen Kompetenzstrukturmodells ein Messinstrument entwickelt, welches das systemische Denken auch bei Grundschülerinnen und Grundschülern, unter Einhaltung psychometrischer Gütekriterien, erfasst. Die Items des Testinstruments wurden anhand der Kennwerte einer ersten Pilotierungsstudie überarbeitet und im Rahmen einer darauffolgenden Hauptstudie an 293 Schülerinnen und Schülern der 4. Jahrgangsstufe der Primarstufe erprobt. Nach Erhöhung des Anforderungsniveaus in der zweiten Erprobung konnten die Itemtrennschärfen in den höheren Dimensionen systemischen Denkens verbessert werden, während die der unteren Dimensionen nach wie vor eher gering ausfallen. Das finale Testinstrument weist eine gute interne Konsistenz auf (Cronbachs  $\alpha = .73$ ). Die Validitätsbetrachtungen anhand eines argumentationsbasierten Ansatzes zeigen u.a. zufriedenstellende Korrelationen systemischen Denkens mit dem Fachwissen zum Klimawandel (konvergente Validität) sowie niedrige Korrelationen mit Personenmerkmalen (diskriminante Validität), was die Konstruktvalidität stützt. Mittels einer konfirmatorischen Faktorenanalyse wurden unterstützende Belege für die strukturelle Validität gesammelt, wobei die Dimensionalität des Konstrukts nur in Teilen nachgewiesen werden konnte. Hinweise auf die Instruktionssensitivität des Messinstruments lieferte der Einsatz in einer Intervention.

**Schlüsselwörter:** Systemisches Denken, Klimawandel, Messinstrument, Grundschule

---



– Original Paper –

---

# The Development and Validation of an Instrument for Measuring Systems Thinking on Climate Change in Elementary School Students

Sven Frey<sup>1</sup>, Maik Beege<sup>2</sup>, Nadine Tramowsky<sup>1</sup> und Werner Rieß<sup>1</sup>

<sup>1</sup>University of Education Freiburg,  
Department of Biology and its Didactics

<sup>2</sup>University of Education Freiburg,  
Department of Psychology

---

## ABSTRACT

In the context of the global challenge of climate change, students should be enabled to understand complex interrelationships as part of education for sustainable development. For this purpose, it is useful to foster their systems thinking, the ability to recognize, describe and model complex aspects of reality as systems. In order to evaluate the effectiveness of instruction in fostering systems thinking, a suitable and manageable diagnostic tool is needed. So far, there is no such tool that captures elementary school students' ability to solve complex problems with system models on climate change. This project developed a test based on a heuristic structural competence model that has proven itself in higher grades. The presented test is designed to assess systems thinking in elementary school students according to psychometric quality criteria. The items of the measuring instrument were revised based on the item parameters of an initial pilot study and tested in a subsequent main study with 293 fourth grade students. After increasing the level of difficulty in the second test, the inter-item correlation in the higher dimensions of systems thinking improved, while that in the lower dimensions was still rather low. The final measuring instrument shows good internal consistency (Cronbach's  $\alpha = .73$ ). Validity analyses using an argumentation-based approach show satisfactory correlations of systems thinking with domain-specific knowledge of climate change (convergent validity) and low correlations with personal characteristics (discriminant validity), which supports construct validity. A confirmatory factor analysis was used to gather supporting evidence in favor of structural validity, although the dimensionality of the construct was only partially supported. Supporting evidence in favor of the instructional sensitivity of the instrument was provided by its use in an intervention.

**Keywords:** systems thinking, climate change, measuring instrument, elementary school

---

## 1 Einleitung

Um mit den Herausforderungen des Klimawandels erfolgreich umzugehen, hat sich die Weltgemeinschaft in der Agenda 2030 auf 17 Nachhaltigkeitsziele (Sustainable Development Goals [SDGs]) verständigt. Diese implizieren unter anderem die Förderung einer *Bildung für nachhaltige Entwicklung (BNE)*, denn es sind vor allem die nachfolgenden Generationen, die mit dem Klimawandel umgehen und diesem entgegenwirken müssen. Die Umsetzung einer BNE findet in Deutschland seit 2017 durch den nationalen Aktionsplan BNE statt (Nationale Plattform Bildung für nachhaltige Entwicklung, 2017) bzw. konkret durch die strukturelle Verankerung der BNE in den Bildungsplänen der Bundesländer. Das Ziel, Schülerinnen und Schüler in einer immer dynamischer und komplexer werdenden Lebenswelt dazu zu befähigen, nachhaltig zu handeln, erfordert eine frühzeitige Förderung von Schlüsselkompetenzen einer BNE, wie etwa systemisches Denken (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2017; Wiek & Redman, 2022). Diese Kompetenz benötigt man vor allem in der Auseinandersetzung mit dem Klimasystem aufgrund dessen hoher systemischer Komplexität. Um die komplexen Strukturen und Prozesse des Klimasystems verständlich zu machen, legt sich die Verwendung von Systemmodellen nahe, welche das mentale Modellieren von Zusammenhängen in der Umwelt unterstützen (bspw. Arndt, 2018; Gesellschaft für Didaktik des Sachunterrichts [GDSU], 2013; Haider, 2019; Matthies, 2010; Ossimitz, 2000). Um zu überprüfen, ob solch eine Förderung erfolgreich ist, sollte die Wirksamkeit unterrichtlicher Fördermaßnahmen mithilfe von Testverfahren evaluiert werden. Nur durch das Zusammenspiel von Diagnose und gezielter Förderung kann eine wirksame BNE erreicht werden. Ein entsprechendes quantitatives deutschsprachiges Testinstrument gibt es für die Zielgruppe der Grundschule zum Thema Klima(wandel) nicht. Bisherige Diagnoseverfahren im systemischen Denken, vor allem im Grundschulkontext, basieren überwiegend auf qualitativen und Mixed-Methods-Ansätzen (Budak & Ceyhan, 2023), welche eine aufwendige Auswertung erfordern. Im Projekt *Systemisches Denken lernen zum Klimawandel (SysDeKlima)* sollte jedoch auf Grundlage des heuristischen Kompetenzstrukturmodells zum systemischen Denken (Rieß, Schuler &

Hörsch, 2015) ein leicht einzusetzendes und auswertbares quantitatives Messinstrument entwickelt werden, welches das systemische Denken im Kontext Klima(wandel) bei Grundschülerinnen und Grundschülern ganzheitlich unter Beachtung psychometrischer Gütekriterien und praktischer Anforderungen erfasst. Ziel der Testentwicklung ist neben der Evaluation von Fördermaßnahmen auch die Erfassung langfristiger Veränderungen, um nachhaltige Effekte unterrichtlicher Maßnahmen zu identifizieren. Darüber hinaus soll die empirische Struktur (*Dimensionalität*) des heuristischen Kompetenzstrukturmodells (Rieß, Schuler & Hörsch, 2015) überprüft werden, die für höhere Jahrgangsstufen bereits gesichert wurde (Brockmüller, 2019), für die Grundschule jedoch noch nicht. Die Testentwicklung orientierte sich an vergleichbaren, empirisch abgesicherten Testinstrumenten (Bräutigam, 2014; Brockmüller, 2019; Fanta, Bräutigam, Greiff & Rieß, 2017), insbesondere hinsichtlich der Itemstruktur. Dabei wurde darauf geachtet, dass die adressierten Kompetenzdimensionen durch hauptsächlich Multiple-Choice-Items adäquat abgebildet werden, während die Anzahl der Items unter Berücksichtigung der Zielgruppe möglichst geringgehalten werden sollte. Allerdings bedurfte es der Anpassung des Kontextes und des kognitiven Niveaus der Lernenden. Beispielsweise konnten Items zum deklarativen systemischen Wissen (z.B. Systemdefinition, Systemstruktur) nach sprachlicher Vereinfachung übernommen werden. Im Gegensatz dazu war die Übernahme oder Anpassung bestehender Instrumente für kontextgebundenen Items mit höheren Kompetenzanforderungen nur in geringem Ausmaß möglich. Folglich musste ein Großteil der Items neu entwickelt werden, um den spezifischen Anforderungen der Grundschule gerecht zu werden. Das Testinstrument soll in einer auf diese Studie folgenden Interventionsstudie eingesetzt werden, um die Wirksamkeit variierter Systemmodelle zum Klimawandel zu erfassen.

## 2 Theoretischer Hintergrund

Unsere schnell verändernde Lebenswelt, unter anderem geprägt von globalem Wandel, technologischem Fortschritt und Umweltveränderungen, stellt Schülerinnen und Schüler vor die Aufgabe, sich in komplexen und sich ständig verändernden Systemen

mit oftmals zeitgleich stattfindenden Prozessen zurechtzufinden. Besonders der Klimawandel erfordert das Verständnis für vielschichtige Wechselwirkungen auf mehreren Systemebenen, welches wiederum Voraussetzung ist, um die eigene Lebenswelt unter dem Ziel der Nachhaltigkeit aktiv mitzugestalten. Dabei stellt sich die Frage, welche spezifischen Kompetenzen erforderlich sind, um diesen komplexen Anforderungen unserer Zeit gerecht zu werden. Die UNESCO (2017) nennt acht Schlüsselkompetenzen zur Erreichung aller SDGs, welche für alle Lernenden aller Altersgruppen weltweit erforderlich sind, darunter das systemische Denken. Mischo und Rieß (2008) definieren dies als kognitive

Fähigkeit, komplexe Wirklichkeitsbereiche als Systeme erkennen, beschreiben und modellieren (z.B. strukturieren, organisieren) zu können. Dazu gehören die Fähigkeiten, Systemelemente und Wechselbeziehungen bestimmen zu können, zeitliche Dimensionen (Dynamiken) erfassen zu können und die Fähigkeit, auf der Basis der eigenen Modellierungen Erklärungen geben, Prognosen treffen und „weiche“ Technologien entwerfen zu können. (S. 348f.)

Unter weichen Technologien versteht man solche, die die emergenten Eigenschaften von Systemen nicht beeinträchtigen. Verhoeff, Knippels, Gilissen und Boersma (2018) betonen, dass das systemische Denken nicht als kohärentes Verstehen in einer Ansammlung von Wissen verstanden werden darf, sondern als Fähigkeit, bei der systemtheoretische Konzepte gezielt zur Erklärung und Vorhersage von Naturphänomenen eingesetzt werden. Betrachtet man das Klima als ein hochkomplexes ökologisches System, bedarf es im Sinne einer grundlegenden systemischen Betrachtung der Identifikation von Systemelementen sowie deren Eigenschaften und Wechselbeziehungen untereinander (Bossel, 1992; Mambrey Timm, Landskron & Schmiemann, 2020). Zudem weist ein System eine Identität auf und steht über die Systemgrenzen hinaus in Beziehung zu seiner Umwelt, welche das Systemverhalten beeinflussen kann. Zustandsänderungen ergeben sich aber auch durch Prozesse innerhalb des Systems, wobei Auswirkungen wiederum auf deren Ursache wirken können. Diese durch äußere oder innere Ursachen bedingten Wirkungsbeziehungen können linear oder nichtlinear sein und je nach Qualität der Wirkung die

Organisation und Funktion des Systems auf unterschiedlichen hierarchischen Ebenen, bis hin zur Bildung neuer Eigenschaften (Emergenz) beeinflussen (Matthies, 2010). Wie können Schülerinnen und Schüler nun befähigt werden, nicht nur die Oberflächenmerkmale eines Systems zu erkennen, sondern auch dessen dynamische Komplexität zu verstehen? Hierzu liefern Ben-Zvi Assaraf und Knippels (2022) etablierte und durch Evidenz gestützte pädagogische Leitlinien und Unterstützungsstrategien zur Förderung systemischen Denkens. Dabei hat sich das Modellieren als besonders zielführende Strategie erwiesen. Die Modellkompetenz steht in engem Zusammenhang mit dem systemischen Denken, insbesondere der Systemmodellierungsfähigkeit, die erforderlich ist, um (komplexe) Systeme durch Modelle adäquat darzustellen. Während die Modellkompetenz (Krüger, Kauertz & Upmeyer zu Belzen, 2018) mit der Kenntnis über Modelle und der Modellbildung vor allem die „Intentionalität von Modellen und ihre Rolle in der Erkenntnisgewinnung fokussiert“ (Terzer, 2012, S. 62), erfordert systemisches Denken darüber hinaus eine ganzheitliche Systembetrachtung, welche auch Dynamiken, Wechselbeziehungen und Rückkopplungen impliziert. Insofern ist die Modellkompetenz als Voraussetzung systemischen Denkens einzuordnen (Terzer, 2012), wobei Schülerinnen und Schüler, insbesondere der Grundschule, in der Bildung mentaler Repräsentationen komplexer Systeme – hier des Klimasystems – unterstützt werden müssen (Acher, 2017; Nitz & Fechner, 2018). Auch das Fachwissen spielt eine zentrale Rolle bei der Förderung systemischen Denkens, da dessen Einbezug zu einem vertieften Verständnis von Ökosystemen führt (Mambrey, Schreiber & Schmiemann, 2020). Menschen ohne inhaltliches Wissen hingegen beschränken sich bei systemischen Betrachtungen eher auf beschreibende und oberflächliche Merkmale (Sweeney & Sterman, 2007). Fachwissen ist daher eine Voraussetzung für ein vertieftes Systemverständnis, reicht jedoch allein nicht aus, um komplexe systemische Probleme zu lösen. Nach der Einordnung und Definition systemischen Denkens werden im Verlauf des Kapitels zunächst unterschiedliche konzeptuelle Ansätze vorgestellt, gefolgt von den entsprechenden Operationalisierungen. Anschließend werden relevante Interventionsstudien diskutiert. Auf Grundlage dieser Ausführungen erfolgt schließlich die theoretische Fundierung dieser Studie.

In der einschlägigen Literatur existieren unterschiedliche Konzeptualisierungen systemischen Denkens (Bielik, Delen, Krell & Ben-Zvi Assaraf, 2023; Budak & Ceyhan, 2023). So koexistieren unterschiedliche Kompetenzmodelle, beispielsweise das Kompetenzmodell von Ossimitz (2000), das Strukturmodell zur Systemkompetenz von Sommer (2005), das System Thinking Hierarchical (STH) Model von Ben-Zvi Assaraf und Orion (2005, 2009), das Structure-Behavior-Function (SBF) Model von Hmelo-Silver und Liu (2007) in Kombination mit dem Components-Mechanisms-Phenomena (CMP) Model von Hmelo-Silver, Jordan, Eberbach und Sinha (2017), das Kompetenzmodell zur Systemkompetenz von Rempfler und Uphues (2011a) und das heuristische Kompetenzstrukturmodell von Rieß et al. (2015). Betrachtet man die unterschiedlichen Konzeptualisierungen systemischen Denkens dieser Kompetenzmodelle, so wird deutlich, dass trotz diverser Unterschiede eine Übereinstimmung in der Mehrdimensionalität des Konstrukts besteht (Mehren, Rempfler, Ullrich-Riedhammer, Buchholz & Hartig, 2016). Folglich ist systemisches Denken keine einzelne isolierbare Fähigkeit, sondern ein Konstrukt, welches mit verschiedenen Kompetenzanforderungen einhergeht. Zudem weisen alle Modelle eine Unterscheidung zwischen einer statischen Systemdarstellung und einer durch Wechselbeziehungen bedingten Interaktion einzelner Systemelemente auf. Im Folgenden wird die unterschiedliche Herangehensweise einiger Autorinnen und Autoren bei der Operationalisierung des systemischen Denkens dargestellt.

Die Messung des systemischen Denkens erfolgt abhängig von der Zielgruppe auf unterschiedliche Weise. Während Ossimitz (2000), Bräutigam (2014), Rempfler und Uphues (2011b) sowie Hmelo-Silver et al. (2017) in der Sekundarstufe, Fanta et al. (2017) bei Lehramtsstudierenden und Brockmüller (2019) in der gymnasialen Oberstufe auf das Testformat Leistungstest (z.T. in Kombination mit Computermodellen) zurückgriffen, erfassen Sommer (2005) sowie Ben-Zvi Assaraf und Orion (2009) das systemische Denken bei Grundschülerinnen und Grundschulern mithilfe umfangreicher Testbatterien. Sommer (2005) nutzte dazu Fragebögen, Befragungen, Begriffslandkarten und einen Computerspieltest, während Ben-Zvi Assaraf und Orion (2009) Schülerzeichnungen, Wortassoziationen,

Interviews, Beobachtungen und Fragebögen einsetzten.

Auf Grundlage der dargestellten Konzeptualisierungen und Operationalisierungen wurden in den verschiedenen Altersgruppen Wirksamkeitsstudien unterschiedlicher Zielsetzungen durchgeführt. Fanta, Bräutigam und Rieß (2020) untersuchten in einer quasi-experimentellen Interventionsstudie das Professionswissen systemischen Denkens bei Lehramtsstudierenden. Weitere quasi-experimentelle Interventionsstudien in der Sekundarstufe untersuchten die Auseinandersetzung mit Modellen in bestimmten Lernarrangements auf unterschiedliche Dimensionen systemischen Denkens (Brockmüller, 2019; Evagorou, Korfiatis, Nicolaou & Constantinou, 2009; Hmelo-Silver et al., 2017; Ossimitz, 2000; Rieß & Mischo, 2009). Für den Primarbereich liegen Studien vor, die das systemische Denken bereits ab der ersten Jahrgangsstufe nachweisen und in quasi-experimentellen Interventionsstudien zu unterschiedlichen Themen durch die Vermittlung von Fachwissen, den Besuch außerschulischer Lernorte, die Durchführung von Experimenten und den Einsatz von Simulationen, intraindividuell oder im Vergleich zur Kontrollgruppe, erfolgreich fördern konnten (Ben-Zvi Assaraf & Orion, 2009; Bertschy, 2008; Sommer, 2005). Darüber hinaus wurde das systemische Denken junger Schülerinnen und Schüler hinsichtlich kognitiver Verstehens- und Argumentationsprozesse untersucht (Hokayem & Gotwals, 2016; Mambrey et al., 2020a; Mambrey et al., 2020b). Studienübergreifend lässt sich feststellen, dass sich der Einsatz von Modellen (analog, digital, Modellierungstechniken) im Unterricht bewährt. Betrachtet man die theoretische Fundierung einiger der genannten Interventionsstudien im Primarbereich, so lassen unterschiedliche Konzeptualisierungen und die Vielfalt der inhaltlich zugrundeliegenden Systeme (bspw. Wasserkreislauf, Storch, Zelle) eine Vergleichbarkeit der Ergebnisse nur schwer zu (Boersma, Waarlo & Klaassen, 2011).

Auf Grundlage der dargestellten Konzeptualisierungen, Operationalisierungen und Wirkungsstudien wird für diese Studie das heuristische Kompetenzstrukturmodell zum systemischen Denken verwendet (Rieß et al., 2015). Zum einen fokussiert es die Modellierung von Systemen, was sich sowohl in der Klimaforschung etabliert (Rieß et al., 2015) als auch in den beschriebenen Wirkungsstudien als lernför-

derlich herausgestellt hat. Zum anderen erfasst dieses Kompetenzstrukturmodell die Ganzheit systemischen Denkens von der Sekundarstufe bis ins Erwachsenenalter und lässt sich leicht an das Niveau der Grundschule adaptieren. Zugunsten einer künftigen Vergleichbarkeit von Studienergebnissen und im Sinne eines ganzheitlichen spiralcurricularen Unterrichts naturwissenschaftlicher Fächer intendieren wir die Tauglichkeit auch für die Grundschule nachzuweisen, sodass sich dieses Kompetenzmodell über alle Niveaustufen etablieren kann. Hierbei ist anzumerken, dass in Anbetracht der Zielgruppe, wie auch bei Fanta et al. (2020), nicht alle Dimensionen beachtet werden können (vgl. Kapitel 4). Zuletzt hat sich das Kompetenzstrukturmodell beim Einsatz in Studien in der Sekundarstufe (Brockmüller, 2019, 2023), bei Lehramtsstudierenden (Fanta et al., 2017; Fanta et al., 2020) sowie in der Lehrerfortbildung (Streiling, Hörsch & Rieß, 2021) als sensitiv erwiesen.

Da hier ein für die Schulpraxis handhabbares Diagnoseinstrument entwickelt werden soll, welches eine zeitökonomische Auswertung erlaubt, wurde bei der Operationalisierung das Format des Tests gewählt. Bereits vorhandene Instrumente enthielten entweder qualitative Elemente (z.B. Erstellung von Concept Maps), welche die Auswertung aufwändig machen, thematisierten andere Inhalte oder waren aufgrund des zu hohen Schwierigkeitsgrades nicht für die Grundschule geeignet. Somit bedarf es einer Neuentwicklung, wobei wir uns an Vorarbeiten von Bräutigam (2014), Brockmüller (2019) und Fanta et al. (2017) orientieren und das systemische Denken unter Beachtung der Domänenspezifität (Klieme & Maichle, 1991) zum Thema Klima(wandel) erfassen.

### 3 Fragestellung

Möchte man Evidenz über die Wirksamkeit von Fördermaßnahmen auf das systemische Denken erhalten, so bedarf es eines Messinstruments, das dieses adäquat erfasst. Zur Untersuchung des Einflusses variierter Systemmodelle zum Klimawandel auf das systemische Denken bei Schülerinnen und Schülern der Grundschule soll ein Messinstrument entwickelt und validiert werden, welches die geförderten Teilfähigkeiten systemischen Denkens präzise misst. Hieraus ergibt sich folgende Fragestellung:

Lässt sich auf Grundlage des in höheren Jahrgangsstufen bewährten heuristischen Kompetenzstrukturmodells ein Messinstrument entwickeln, welches das systemische Denken auch bei Grundschülerinnen und Grundschülern reliabel und valide erfasst? Um diese Frage zu beantworten, werden die folgenden zwei Schritte unternommen: In einer ersten Pilotierungsstudie sollen erste Items entwickelt und erprobt werden. Nach der Eliminierung oder Überarbeitung psychometrisch kritischer Items werden diese nach einer zweiten Studie erneut auf Grundlage der Kennwerte evaluiert. Zusätzlich werden durch die Sammlung empirischer Evidenz Validitätsbetrachtungen vorgenommen, um die Testwerte hinsichtlich der beabsichtigten Verwendung des Tests zu interpretieren.

*Validitätskonzept.* Nach der American Educational Research Association (AERA) bezieht sich Validität auf das Ausmaß “to which evidence and theory support the interpretation of test scores for proposed uses of tests” (AERA, 2014, S. 11). Validität wird also nicht als Eigenschaft eines Tests verstanden, sondern als Kriterium hinsichtlich der Zulässigkeit von Testwertinterpretationen, welche wiederum abhängig von der beabsichtigten Verwendung des Tests sind (Hartig, Frey & Jude, 2020). Wir intendieren mit dem zu entwickelnden Testinstrument die Dimensionen systemischen Denkens zum Klimawandel in der vierten Jahrgangsstufe zu messen, um Interventionsmaßnahmen zu evaluieren und Lernentwicklungen darstellen zu können. Vor diesem Ziel und dem argumentationsbasierten Validitätskonzept von Kane (2013) folgend, werden in der Hauptstudie dieses Artikels zur Prüfung der Validität zum einen unterstützende Belege in Form empirischer Evidenzen für die Konstruktvalidität (Beziehung des Konstrukts zu anderen Variablen) gesammelt, indem im Sinne konvergenter Validität der korrelative Zusammenhang mit einem theoretisch verwandten Konstrukt (Fachwissen) (Schmidt-Atzert & Amelang, 2012) und im Sinne diskriminanter Validität der korrelative Zusammenhang mit unterschiedlichen Konstrukten (Personenmerkmale) (Moosbrugger & Kelava, 2020) untersucht wird. Zum anderen werden Hinweise für die strukturelle Validität (Dimensionalität) mittels konfirmatorischer Faktorenanalyse (KFA) gesammelt. Um dem Ziel der Evaluation von Interventionsmaßnahmen gerecht zu werden, braucht es zudem Belege für die

Instruktionssensitivität, welche durch den praktischen Einsatz des Tests in einer Intervention gewonnen werden sollen. Die Gesamtheit der hier dargestellten empirischen Evidenzen soll letztlich das Validitätsargument stützen.

## 4 Methode

*Testkonstruktion.* Um Testwertinterpretationen zuzulassen, bedarf es vorab der genauen Definition des zu messenden Konstrukts (Hartig et al., 2020). Dieses wird im Folgenden dargelegt. Der Test wurde auf Grundlage des heuristischen Kompetenzstrukturmodells (Rieß et al., 2015) erstellt (Kapitel 2). Aus verschiedenen Gründen entschieden wir uns, bei der Entwicklung des Testinstruments nicht alle Teilfähigkeiten zu berücksichtigen. Zum einen wäre ein Test, der alle Teilfähigkeiten erfasst, für die Grundschule zu umfangreich; zudem musste das Fachwissen in einem separaten Test erfasst werden, da dieses als ein bedeutender Prädiktor systemischen Denkens gilt (Sommer, 2005; Mambrey et al., 2020a) und somit in der Datenauswertung berücksichtigt werden

sollte. Zum anderen lag ein besonderer Fokus auf Kompetenzdimension 3 (Lösen komplexer Probleme mit Hilfe von Systemmodellen), da der Einsatz von Systemmodellen als übergreifende pädagogische Leitlinie zur effektiven Förderung systemischen Denkens gilt (Ben-Zvi Assaraf & Knippels, 2022). Ungeklärt ist jedoch, welche Art von Systemmodell (qualitativ oder quantitativ) für den Bereich der Primarstufe besonders wirksam ist. Das in dieser Studie entwickelte Testinstrument soll in unserer Folgestudie eingesetzt werden, um die Wirksamkeit variiert Systemmodelle zum Klimawandel zu erfassen. Ausgehend davon wurden in den Dimensionen 1 (Deklaratives/konzeptuelles systemisches Wissen) und 2 (Systemmodellierungsfähigkeit) nur die Teilfähigkeiten im Testinstrument berücksichtigt, die als Vorläuferfähigkeiten für die Zielfähigkeiten in Dimension 3 zu sehen sind. Aus den genannten Gründen haben wir die in Tabelle 1 dargestellten sechs Teilfähigkeiten selektiert und diese – wo nötig – an das Niveau der vierten Jahrgangsstufe der Grundschule angepasst. Entsprechende Anpassungen werden im Folgenden kurz erläutert.

Tabelle 1

Ausgewählte und an das Niveau der Grundschule angepasste Kompetenzen systemischen Denkens (in Anlehnung an das heuristische Kompetenzstrukturmodell von Rieß et al., 2015)

<b>Dimension 3: Lösen komplexer Probleme mit Hilfe von Systemmodellen</b>	Mit Hilfe qualitativer Systemmodelle Erklärungen geben, Prognosen treffen und Technologien entwerfen können	Mit Hilfe quantitativer Systemmodelle Erklärungen geben, Prognosen treffen und Technologien entwerfen können
<b>Dimension 2: Systemmodellierungsfähigkeit</b>	Systemisch modellierbare Wirklichkeitsbereiche mit Hilfe eines Textes / Wortmodells verstehen können	Qualitative Systemmodelle lesen können / Vernetzungskreis verstehen können
<b>Dimension 1: Deklaratives/ konzeptuelles systemisches Wissen</b>	Systemtheoretisches Grundwissen (Systembegriff, Systemstruktur, Systemverhalten)	Kenntnis von exemplarischen Systemen
<b>Kompetenzdimensionen</b>	<b>Teilfähigkeit 1</b>	<b>Teilfähigkeit 2</b>

Die erste Kompetenzdimension ist durch deklaratives/konzeptuelles systemisches Wissen gekennzeichnet. Im Testinstrument werden neben dem Systembegriff „Grundsätze einer systemwissenschaftlichen Betrachtungsweise von Ausschnitten der Wirklichkeit“ (Rosenkränzer, Stahl, Hörsch, Schuler & Rieß, 2016, S. 110) mit exemplarischer Kenntnis erfasst. Angesichts der Zielgruppe ist anzumerken, dass es sich hierbei um ein basales Verständnis systemischen Denkens handelt. In einer ersten Ausei-

nersetzung mit Systemen können zugunsten eines grundlegenden Verständnisses Eigenschaften komplexer Systeme, wie zum Beispiel Multikausalität, Rückkopplungen, dynamische Komplexität (Osimitz, 2000), Nichtlinearität, Emergenzen oder dissipative Strukturen (Rosenkränzer et al., 2016) nur bedingt thematisiert werden.

Die zweite Kompetenzdimension umfasst die Systemmodellierungsfähigkeit und somit die Fähigkeit Systemmodelle verschiedener Art verstehen und

konstruieren zu können. An dieser Stelle wurden in Anlehnung an das heuristische Kompetenzstrukturmodell (Rieß et al., 2015) Anpassungen dahingehend vorgenommen, dass für Schülerinnen und Schüler der Grundschule vorerst das *Verstehen* gefördert werden sollte, bevor es zur selbstständigen Konstruktion von Systemmodellen kommt. So sollen, beginnend mit bekannten Repräsentationsformen (beispielsweise Text) über die Reduktion von Merkmalen (beispielsweise Wortmodell) immer abstraktere Systemmodelle (beispielsweise Vernetzungskreis) verstanden werden können. An dieser Stelle wurde die Methode des Wirkungsgraphen mit der des Vernetzungskreises ersetzt, da sich diese in der Grundschule bereits etabliert hat (Bollmann-Zuberbühler, Frischknecht-Tobler, Kunz, Nagel & Wilhelm Hamiti, 2010). Ein besonderer Vorteil gegenüber des Wirkungsgraphen liegt in der Vorgabe der äußeren Struktur, welche die Systemmodellierung vereinfacht.

*Die dritte Kompetenzdimension* – das Lösen komplexer Probleme mit Hilfe von Systemmodellen – stellt die zentrale Dimension systemischen Denkens mit Systemmodellen zum Klimawandel und somit des Testinstruments dar. Mit Hilfe qualitativer Systemmodelle (Teilfähigkeit 1) und quantitativer Systemmodelle (Teilfähigkeit 2) sollen einfache Erklärungen für Systemverhalten gegeben, Prognosen getroffen und Technologien entworfen werden.

Da systemisches Denken zweifellos eine Vielzahl kognitiver Fähigkeiten umfasst, wurde die Form eines Leistungstests ausgewählt (Moosbrugger & Kelava, 2020). Zugunsten einer ökonomischen Testauswertung zogen wir ein digitales Testformat in Betracht, entschieden uns jedoch dagegen, da in der Primarstufe nicht von einer hinreichenden Medienkompetenz ausgegangen werden kann. Stattdessen wurde der Test in einem Paper-Pencil-Format erstellt, das den Grundschülerinnen und Grundschülern vertraut ist. Zudem wurden überwiegend Items im Multiple-Choice-Format erstellt, um eine kurze Bearbeitungsdauer sowie eine schnelle und objektive Auswertung zu gewährleisten. Zwar gilt die Erstellung von Concept-Maps als bewährtes Verfahren zur Diagnose systemischen Denkens (bspw. Ben-Zvi Assaraf & Knippels, 2022; Sommer 2005), doch würde dies bei der angestrebten großen Stichprobengröße sowohl der Testökonomie als auch der inhaltlichen Ausrichtung des adaptierten Kompetenzmo-

dells widersprechen. Hierbei stehen Verstehensprozesse im Vordergrund, nicht die eigenständige Produktion von Systemmodellen (Concept-Maps/Vernetzungskreise), welche wiederum methodische Vorkenntnisse und Erfahrung erfordern würde (Kinchin, 2020), die zum Zeitpunkt des Prätests nicht vorausgesetzt werden können. Die Testlänge (Anzahl der Testitems) sollte für die hiesige Zielgruppe der Grundschule bewusst geringgehalten werden, um ein Absinken der Arbeitsbereitschaft und der Konzentration zu verhindern.

*Item-Entwicklung.* Bei der Item-Entwicklung orientierten wir uns an den methodischen Empfehlungen von Carpenter (2018) und DeVellis und Thorpe (2022). Diesen folgend analysierten wir bereits bestehende Messinstrumente von Bräutigam (2014), Brockmüller (2019) und Fanta et al. (2017). Diese erfassen das systemische Denken auf Grundlage desselben Kompetenzmodells, jedoch zu anderen Themenbereichen (Ökosystem Wald, Bodenerosion) und in höheren Jahrgangsstufen (Sekundarstufe 1, gymnasiale Oberstufe, Lehramtsstudierende). Einzelne Items (bspw. „Wie denkst du, könnte man allgemein ein System beschreiben?“) konnten nach Anpassung der Formulierung und des Anforderungsniveaus der Antwortoptionen übernommen werden. Andere Items wurden thematisch angepasst, da das systemische Denken entsprechend den Empfehlungen von Klieme und Maichle (1991) domänenspezifisch (hier: Klima[wandel]) erfasst wird. Um sowohl die Passung des thematischen Anforderungsniveaus als auch die Plausibilität aller Antwortoptionen zu gewährleisten, wurden neben der Beachtung relevanter Literatur (Lüschen, 2015) in zwei vierten Klassen der Primarstufe in Baden-Württemberg mittels Wortassoziationen und Fragebögen im offenen Format (bspw. „Was passiert, wenn sich die Temperatur auf der Erde erhöht?“) explorativ Präkonzepte erhoben. Bei der Item-Entwicklung wurde darauf geachtet, dass alle notwendigen Informationen im Itemstamm enthalten sind, sodass keine Messung von Fachwissen, sondern von Kompetenzen stattfindet (Ropohl, Walpuski & Sumfleth, 2015). Zunächst wurde ein Katalog aus 10 Aufgaben (Items) erstellt, der die Dimensionen des Kompetenzmodells abbildet (Abb. 1). Die Verteilung der Items auf die Dimensionen, aufgeführt mit beispielhaften Itemstämmen für die jeweiligen Teilfähigkeiten, zeigt sich wie folgt:

- drei Aufgaben für Dimension 1 (bspw. „Was denkst du, wie könnte man allgemein ein System beschreiben?“ [Teilfähigkeit 1]; „Bei folgenden Bereichen handelt es sich jeweils um ein System: das Klima, der Computer, der Körper. Schreibe 4 weitere Systeme auf.“ [Teilfähigkeit 2])
- drei Aufgaben für Dimension 2 (bspw. „Welches der folgenden Pfeilbilder passt zu den Informationen aus dem Text?“ [Teilfähigkeit 1]; „Welche der unteren Aussagen passen zum Vernetzungskreis?“ [Teilfähigkeit 2])
- vier Aufgaben für Dimension 3 (bspw. „Wie wirkt sich der Klimawandel mit der Zeit auf den Meeresspiegel aus? Verbinde.“ [Teilfähigkeiten 1-2])

Insgesamt besteht der Test aus acht geschlossenen Aufgaben im Multiple-Choice-Format (unter mehreren Antworten auswählen), einer im halboffenen Format (Einwortantworten) und einer im offenen Format (Freitextantwort). Die geringe Itemanzahl erfordert die Messung des systemischen Denkens auf Ebene der Dimensionen, auch wenn diese nicht unidimensional sind und aus mehreren Teilfähigkeiten bestehen.

*Prozedur und Material.* Die Datenerhebung zur Itemanalyse fand während des regulären Unterrichts statt. Hierzu wurde der Test zum systemischen Denken zum Klimawandel von allen Schülerinnen und Schülern zunächst als Baseline-Messung im Paper-Pencil-Format bearbeitet, da sie mit diesem Testformat vertraut sind. Die Testung wurde von der jeweiligen Klassenlehrkraft durchgeführt, welche zur Gewährleistung der Durchführungsobjektivität zuvor genau instruiert wurde und einem exakten Testmanual folgte. Nach einer Intervention im Umfang von fünf Unterrichtsstunden je 45 Minuten entsprechend dem Modell problemorientierten Lehrens und Lernens (Rieß & Mischo, 2017) zur Förderung systemischen Denkens zum Klimawandel wurde dieser Test sieben Tage später noch einmal wiederholt, um Argumente für die Instruktionssensitivität als Aspekt der Validität zu sammeln. Ziel war es, zu untersuchen, ob der Test nicht nur als Erhebungsinstrument für systemisches Denken dient, sondern auch sensitiv auf Veränderungen durch Interventionen reagiert.

*Testauswertung.* Bei den Aufgaben im Multiple-Choice-Format wurde für jede Antwortmöglichkeit

ein Punkt vergeben, da immer aufs Neue eine Entscheidung zwischen richtig (ankreuzen) und falsch (nicht ankreuzen) gefällt werden musste. In den übrigen Aufgabenformaten (offen, halboffen, geschlossen) wurde für die jeweils richtige Antwort oder Erklärung ein Punkt vergeben. Basierend auf dieser Punktevergabe wurde zur Gewährleistung der Auswertungsobjektivität vor der Testauswertung eine Bewertungsmatrix für offene und halboffene Fragen erstellt. Das weitere Vorgehen wird für die jeweiligen Studien spezifiziert dargestellt.

## 5 Pilotierungsstudie: Itemanalyse und Itemselektion

Ziel der Pilotierungsstudie war es, die Testitems hinsichtlich ihrer Schwierigkeit, Varianz und Trennschärfe zu prüfen, um deren Eignung zur Erhebung des Konstrukts sicherzustellen. Der Test wurde sowohl vor als auch nach einer fünfstündigen Intervention zur Förderung systemischen Denkens mit Systemmodellen zum Klimawandel durchgeführt. Dabei wurden Interventionseffekte untersucht, um festzustellen, ob der Test sensitiv auf Veränderungen im systemischen Denken reagiert. Der Fokus der Pilotierungsstudie lag somit auf der Qualität und Sensitivität der Items. Eine umfassende, auf empirischer Evidenz basierende Validitätsprüfung zur Stützung der Interpretation und Anwendung des Tests erfolgte in der späteren Hauptstudie.

*Stichprobe.* Die Erhebung fand mit 43 Schülerinnen und Schülern der vierten Jahrgangsstufe der Primarstufe in Baden-Württemberg statt.

### 5.1 Analyseplan

*Scoring.* In der Pilotierungsversion des Testinstruments konnte eine Maximalpunktzahl von 42 Punkten (17 Punkte in Dimension 1, 8 Punkte in Dimension 2 und 17 Punkte in Dimension 3) erreicht werden. Dabei konnte für jede zu treffende Entscheidung ein Punkt erzielt werden. In Dimension 2 waren weniger Punkte zu erreichen, da diese Dimension aufgrund von Anpassungen im Kompetenzstrukturmodell an die Zielgruppe der Primarstufe mit weniger Items repräsentiert wurde (vgl. Kapitel 4).

*Berechnungen.* In der Pilotierungsstudie wurden sämtliche Aufgaben anhand einer Itemanalyse (Itemschwierigkeit, Varianz, Itemtrennschärfen, Reliabilitäten) untersucht und selektiert. Ziel war die

Auswahl geeigneter Aufgaben und die Identifizierung von Aufgaben, die sich als nicht adäquat erwiesen. Die Analysen fanden mittels *IBM SPSS Statistics* statt (IBM Corporation, 2021). Hierbei wurden sämtliche Dimensionen zunächst separat untersucht, um unterschiedliche Wirkungen von Interventionen mit Systemmodellen zum Klimawandel sowohl allgemein als auch differenziert betrachten zu können. Die korrespondierenden Aufgaben der jeweiligen Dimension wurden nach Kelava und Moosbrugger (2020) deskriptiv analysiert. Das heißt, es erfolgte zunächst die Auswertung der Itemschwierigkeiten, um auszuschließen, dass Items entweder zu leicht oder zu schwierig waren oder nicht sensitiv auf Interventionen reagierten. Aufgaben mit mittlerer Schwierigkeit bei der Nachtestung wurden beibehalten, um das Risiko von Decken- oder Bodeneffekten zu minimieren. Zu anspruchsvolle oder zu einfache Aufgaben wurden entsprechend angepasst oder entfernt (Bortz & Döring, 2006). Zusätzlich wurden die Itemtrennschärfen für die Kompetenzdimensionen berechnet, da diese ein Hauptkriterium zur Aufgabenselektion sind (McDonald, 1999). Aufgaben, welche eine geringe Trennschärfe aufwiesen, wurden überarbeitet oder entfernt (<0.3, vgl. Weise, 1975). Danach erfolgte die Berechnung der Reliabi-

lität. Um die interne Konsistenz der Dimensionen zu überprüfen, wurde Cronbachs Alpha ( $\alpha$ ; Cronbach, 1951) berechnet. Um zu testen, ob der Test sensitiv auf die Intervention reagiert, wurden Summenscores für jede Dimension berechnet (vor und nach der Intervention) und mittels t-Tests für Messwiederholungen Unterschiede (Wissenszuwächse) spezifiziert. Angesichts der praktischen Erprobung des Testinstruments im Rahmen einer Intervention unter real schulischen Bedingungen entschieden wir uns gegen die Datenauswertung mittels IRT (Item-Response-Theorie). Unter dem Ziel ein für die Schulpraxis handhabbares Testinstrument zu entwickeln, legten wir mehr Wert auf Erkenntnisse über die Anwendbarkeit und Praktikabilität des Instruments als auf die komplexeren Modellierungen der IRT. Zudem erschien uns, aufgrund der begrenzten Stichprobengröße in der Pilotierungsstudie, der Einsatz klassischer Analyseverfahren als angemessen und ausreichend.

## 5.2 Ergebnisse

Die deskriptiven Itemcharakteristika sind in Tabelle 2 dargestellt. Sie zeigt die Itemkennwerte für den Test zum systemischen Denken vor der Intervention ( $t1$ ) und nach der Intervention ( $t2$ ).

Tabelle 2

*Pilotierungsstudie: Itemkennwerte vor der Intervention ( $t1$ ) und nach der Intervention ( $t2$ )*

Dimension /Item	Itemschwierigkeit	Varianz	Trennschärfe
<b>Dimension 1 (Deklaratives/konzeptuelles systemisches Wissen)</b>			
Item 1 ( $t1$ )	.89	.07	.01
Item 1 ( $t2$ )	.98	.01	.28
Item 2 ( $t1$ )	.86	.12	.09
Item 2 ( $t2$ )	.86	.13	.35
Item 3 ( $t1$ )	.93	.03	-.03
Item 3 ( $t2$ )	.97	.01	.55
Item 4 ( $t1$ )	.54	.03	-.05
Item 4 ( $t2$ )	.88	.03	.35
<b>Dimension 2 (Systemmodellierungsfähigkeit)</b>			
Item 5 ( $t1$ )	.85	.10	.05
Item 5 ( $t2$ )	.88	.06	.18
Item 6 ( $t1$ )	.81	.08	.05
Item 6 ( $t2$ )	.92	.01	.18

Dimension /Item	Itemschwierigkeit	Varianz	Trennschärfe
<b>Dimension 3 (Lösen komplexer Probleme)</b>			
Item 7 (t1)	.75	.16	.02
Item 7 (t2)	.85	.13	.39
Item 8 (t1)	.67	.01	.18
Item 8 (t2)	.75	.02	.44
Item 9 (t1)	.71	.11	.49
Item 9 (t2)	.79	.04	.36
Item 10 (t1)	.40	.25	.26
Item 10 (t2)	.43	.25	.33

*Test vor der Intervention (t1).* Die Kennwerte zeigen, dass vor allem die Items der ersten beiden Dimensionen deutlich zu leicht sind. Diese werden auch ohne Intervention nahezu von allen Schülerinnen und Schülern korrekt beantwortet. In der dritten Dimension sind die Schwierigkeiten akzeptabler, aber ebenfalls teilweise über .70. Mit Ausnahme eines Items in der dritten Dimension sind die Trennschärfen durchgehend ungenügend, was sich auch in der geringen Reliabilität der Dimensionen niederschlägt: Dimension 1:  $\alpha = -.01$ , Dimension 2:  $\alpha = .09$ , Dimension 3:  $\alpha = .35$  (gesamter Test:  $\alpha = .60$ ).

*Test nach der Intervention (t2).* Die Kennwerte zeigen, dass die Itemschwierigkeiten nach der Intervention noch geringer werden. Nun werden innerhalb der ersten beiden Dimensionen Deckeneffekte festgestellt. Allerdings zeigt sich, dass die Itemtrennschärfen deutlich besser ausfallen. Die Reliabilitäten sind höher als beim Test vor der Intervention, jedoch noch nicht im zufriedenstellenden Bereich: Dimension 1:  $\alpha = .46$ , Dimension 2:  $\alpha = .30$ , Dimension 3:  $\alpha = .56$  (gesamter Test:  $\alpha = .63$ ).

*Interventionseffekt.* t-Tests zeigten, dass sich die Schülerinnen und Schüler in jeder Dimension steigern konnten. Hierbei wurden mittlere bis hohe Effektstärken beobachtet: Dimension 1:  $t(41) = 6.09$ ,  $p < .001$ ,  $d = 0.94$ , Dimension 2:  $t(41) = 3.15$ ,  $p = .002$ ,  $d = 0.49$ , Dimension 3:  $t(41) = 2.96$ ,  $p = .003$ ,  $d = 0.46$  (gesamter Test:  $t(41) = 6.47$ ,  $p < .001$ ,  $d = 1.00$ ).

### 5.3 Diskussion

Der bisherige Aufgabenpool war zu leicht, weshalb das Anforderungsniveau grundsätzlich vor der Hauptstudie erhöht wurde. Hierbei ist anzumerken, dass die Teilfähigkeiten in den Kompetenzdimensionen 1 und 2 als Vorläuferfähigkeiten zu betrachten

sind, welche Voraussetzung für das komplexe Problemlösen in Dimension 3 sind. Daher ist eine zumindest halbwegs geringe Schwierigkeit in diesen ersten beiden Dimensionen zu erwarten. Dies schlägt sich ebenfalls in Trennschärfen und Reliabilitäten nieder. Weiterhin gehen die geringen Reliabilitäten und Trennschärfen ebenfalls auf das zugrunde liegende heuristische Kompetenzstrukturmodell (Rieß et al. 2015) zurück, da die Dimensionen nochmals in Teilfähigkeiten unterteilt sind. Dadurch ist ein Maß der internen Konsistenz nur eingeschränkt interpretierbar. Aufgrund der Eigenschaften des Konstrukts und der Empfehlung von Stadler, Sailer und Fischer (2021) folgend, soll und wird die Itemselektion nicht unter dem Gesichtspunkt der Maximierung der internen Konsistenz durchgeführt. Vielmehr wird auf theoretischer Ebene argumentiert, welche Items essenziell sind, um die jeweilige Kompetenzdimension inhaltlich adäquat zu repräsentieren. Weiterhin wird in der folgenden Itemselektion und -adaption darauf abgezielt alle Dimensionen inhaltlich zu erhalten. Besonderes Augenmerk wird dabei erneut auf Dimension 3 gelegt, da der Test in einer Folgestudie eingesetzt werden soll, um die Wirksamkeit von qualitativen und quantitativen Systemmodellen zu überprüfen. Eine Stärke der vorliegenden ersten Version der Skala ist die Sensitivität auf instruktionale Interventionen. Jedoch sind die Interventionseffekte aufgrund der geringen Reliabilitäten vorsichtig zu interpretieren.

## 6 Hauptstudie: Itemanalyse und Validitätsbetrachtungen

*Testüberarbeitung.* Die erste Pilotierung machte deutlich, dass viele Items überarbeitet oder entfernt und komplett neu erstellt werden müssen. Ziel der

Hauptstudie ist somit, die Kennwerte der überarbeiteten Items erneut darzustellen und unterstützende Belege für die intendierten Testwertinterpretationen zu sammeln, um das Validitätsargument zu stützen (Hartig, Frey & Jude, 2020; Kane, 2013). Damit soll sichergestellt werden, dass die Testwerte in Bezug auf die beabsichtigte Verwendung des überarbeiteten Tests sinnvoll interpretiert werden können. In den Kompetenzdimensionen 1 und 2 stellten sich einige Items als zu einfach heraus. Dementsprechend konnte das generelle Niveau angehoben werden, etwa durch das Hinzufügen von Antwortalternativen und Wechselwirkungen, die Veränderung des Kontextes und die Erhöhung der Komplexität von Antwortmöglichkeiten.

*Dimension 1 (Deklaratives/konzeptuelles systemisches Wissen).* In der ersten Dimension erfassen nun die Aufgaben 1-3 das systemtheoretische Grundwissen (Aufgabe 1: Was ist ein System?; Aufgabe 2: Dynamik von Systemen; Aufgabe 3: Kenntnis exemplarischer Systeme). *Dimension 2 (Systemmodellierungsfähigkeit).* In der zweiten Dimension erfordert Aufgabe 4 die Identifikation eines zum Text passenden Wortmodells, während in Aufgabe 5 korrekte Aussagen zu einem Vernetzungskreis ausgewählt werden müssen. *Dimension 3 (Lösen komplexer Probleme mit Hilfe von Systemmodellen).* Aufgrund der inhaltlichen Bedeutsamkeit von Kompetenzdimension 3, wurden die Teilfähigkeiten dieser Dimension im überarbeiteten Test deutlich differenzierter erfasst. Daher wurden hier zu den bestehenden und überarbeiteten Items vier neue erstellt. Die Fokussierung dieser Dimension durch eine erhöhte Anzahl von Items ist notwendig, um in der Folgestudie umfassende Aussagen über die Wirksamkeit der verschiedenen Systemmodelle zum komplexen Problemlösen treffen zu können. Deshalb wurden die Teilfähigkeiten der dritten Dimension nochmals unterteilt in *Erklärungen* (Aufgaben 6-8), *Prognosen* (Aufgaben 9-11), und *Technologien* (Aufgabe 12). So erfasst die neu erstellte Aufgabe 6 das Geben von Erklärungen mithilfe eines qualitativen Systemmodells, die Aufgabe 7 mithilfe eines quantitativen Systemmodells. In Aufgabe 8 wird überprüft, ob auf Grundlage der Erkenntnisse aus den Aufgaben 6 und 7 nun allgemeine Erklärungen zum Klimawandel generiert werden können. In den Aufgaben 9-11 gilt es Prognosen zu erstellen, indem spezifische Veränderungen innerhalb der Teilsysteme (Aufgabe 9), aber auch globale Auswirkungen auf die Temperatur

identifiziert werden sollen (Aufgabe 10). In Aufgabe 11 wird erfasst, ob indirekte globale Ursache-Wirkungsbeziehungen verstanden werden und ein sich verstärkender Rückkopplungsprozess erkannt werden kann. In Aufgabe 12 sollen auf dem Niveau der Primarstufe weiche Technologien entworfen werden, also solche, die Systemdynamiken nicht ändern und Systemeigenschaften sowie das System selbst nicht zerstören (Mischo & Rieß, 2008). Diese Fähigkeit wird mit nur einer Aufgabe erfasst, da es ausreichend die zentrale Ursache des Klimawandels – der Anstieg von Treibhausgasen in der Atmosphäre – auszumachen, um darauf basierend eine Erfindung zu entwerfen, die dieses Problem löst. Hier wurde beispielsweise oft ein CO<sub>2</sub>-Staubsauger genannt. Weitere Items für die Teilfähigkeit Technologie wurden unter Betracht der Zielgruppe der Grundschule nicht erstellt, da diese eher die Kreativität als die tatsächlich dahinterliegende Ursache messen würden. Im Folgenden sind beispielhafte Items für die einzelnen Dimensionen dargestellt.

Was ist ein System?  
Kreuze eine Antwort an.

Ein System besteht aus mehreren Teilen. Diese Teile haben nichts miteinander zu tun.

Ein System besteht aus mehreren Teilen. Diese Teile sind miteinander verbunden.

Ein System besteht aus höchstens zehn Teilen. Diese Teile sind miteinander verbunden.

Ein System besteht aus mehreren Teilen. Diese Teile sind miteinander verbunden und werden immer mehr und mehr.

Abbildung 1. Beispielitem aus Dimension 1 (Systemtheoretisches Grundwissen)

Welche der unteren Aussagen passen zum Vernetzungskreis?  
Kreuze an. Die Kästen helfen dir dabei.

+	Wenn mehr ..., dann mehr ...	Wenn mehr Autos, dann mehr Treibhausgase.
	Wenn weniger ..., dann weniger ...	Wenn weniger Autos, dann weniger Treibhausgase.
-	Wenn mehr ..., dann weniger ...	Wenn mehr Gletscher, dann weniger Meeresspiegel.
	Wenn weniger ..., dann mehr ...	Wenn weniger Gletscher, dann mehr Meeresspiegel.

Wenn es weniger Treibhausgase gibt, dann wird es auf der Erde wärmer.

Wenn mehr Menschen Fahrradfahren, dann sinkt der Meeresspiegel.

Wenn die Gletscher schmelzen, dann steigt die Temperatur (es wird wärmer).

Wenn es mehr Rinder und Kühe gibt, dann sinkt die Temperatur (es wird kälter).

Wenn weniger Menschen Autofahren, dann gibt es mehr Gletscher.

Abbildung 2. Beispielitem aus Dimension 2 (Qualitative Systemmodelle lesen können)

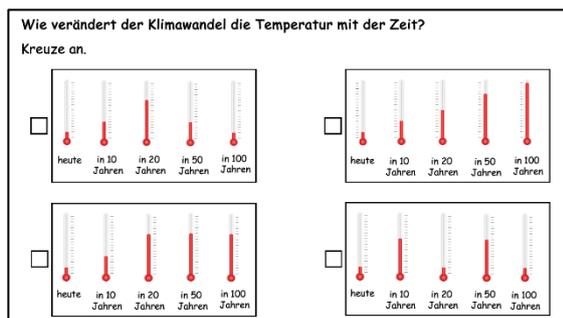


Abbildung 3. Beispielimitem aus Dimension 3 (Prognosen treffen)

*Stichprobe.* Die Erhebung fand mit 293 Schülerinnen und Schülern (56% weiblich) der vierten Jahrgangsstufe der Primarstufe in Baden-Württemberg statt.

### 6.1 Analyseplan

*Scoring.* In der finalen Testversion konnte eine Maximalpunktzahl von 40,5 Punkten erreicht werden. Zur Wahrung der Interpretationsobjektivität wurde die Auswertung von drei Personen vorgenommen, die alle über einen fachlichen Hintergrund im systemischen Denken verfügten und vor der Testauswertung ein Auswertungstraining absolvierten. Die Interraterkorrelation zwischen den drei unabhängigen Ratern wurde für jedes Item separat mithilfe des *Intraclass Correlation Coefficient* (ICC) für 10% der Fälle berechnet. Die Werte lagen zwischen .87 und .99, was auf eine gute Übereinstimmung zwischen den Ratern hinweist (Koo & Li, 2016).

*Berechnungen.* Die Itemanalyse verlief äquivalent zur Pilotierungsstudie. Zusätzlich wurde zur Überprüfung der strukturellen Validität eine *konfirmatorische Faktorenanalyse* (KFA) mittels JASP Version 0.19.1 durchgeführt. Die KFA ermöglicht es, die Struktur latenter Variablen zu testen und zu validieren, indem sie prüft, ob die empirischen Daten mit dem zugrundeliegenden heuristischen Kompetenzstrukturmodell (Rieß et al., 2015) übereinstimmen. Dabei wurden die Ladungen der Items auf die jeweiligen Faktoren (Dimensionen) getestet, um sicherzustellen, dass die verwendeten Items die entsprechenden Kompetenzdimensionen repräsentieren. Zur Beurteilung der Modellpassung wurden der *Comparative Fit Index* (CFI) und der *Root Mean Square Error of Approximation* (RMSEA) für die jeweiligen Messzeitpunkte angegeben. CFI-Werte nahe 0.95 (Hu & Bentler, 1999) und RMSEA-Werte

unter 0.10 (Browne & Cudeck, 1992) wurden als akzeptabel angesehen. Weitere Validitätsbetrachtungen erfolgten durch eine umfassende Untersuchung der Konstruktvalidität (Carlson & Herdman, 2021). Die Frage hierbei ist, wie der Test in das nomologische Netzwerk mit anderen theoretisch verwandten Konstrukten passt (Borsboom, Mellenbergh & van Heerden, 2004; Colliver, Conlee & Verhulst, 2012). Die Quantifizierung der Konstruktvalidität eines Instruments erfolgt größtenteils durch die Identifizierung von Korrelationen mit verschiedenen Maßen. Die resultierenden Korrelationsmuster liefern Informationen über den Grad der Übereinstimmung zwischen dem Maß und den theoretisch vorhersagbaren Variablen (Westen & Rosenthal, 2003). Basierend auf Campbell und Fiske (1959) kann die Konstruktvalidität in konvergente und diskriminante Validität unterteilt werden. Konvergente Validität liegt vor, wenn verschiedene Instrumente, die dasselbe oder ein verwandtes Konstrukt messen sollen, stark miteinander korrelieren (Schmidt-Atzert & Amelang, 2012). Im Gegensatz dazu wird von diskriminanter Validität gesprochen, wenn Instrumente, die unterschiedliche Konstrukte messen, nicht oder nur schwach miteinander korrelieren (Moosbrugger & Kelava, 2020). Zur Berechnung der konvergenten Validität wurde ein Test zum Fachwissen zum Klima(wandel) erstellt, da in der geplanten und auf ihre Wirksamkeit zu überprüfende Intervention der Klimawandel Thema und Inhalt des Unterrichts sein soll. Dieser Test wurde für das Anspruchsniveau der Primarstufe erstellt und ebenfalls vor und nach der Intervention eingesetzt. Er bestand aus 9 Fragen ( $t1$ :  $\alpha = .61$ ;  $t2$ :  $\alpha = .66$ ) im Multiple-Choice-Format und wurde von Experten der Biologie und Geographie hinsichtlich fachlicher Korrektheit, inhaltlicher Validität, Anforderungsniveau und der beabsichtigten Verwendung evaluiert. Inhaltlich erfasste der Test fachliche Grundlagen zum Klima(wandel) (bspw. Klima in Abgrenzung zum Wetter) sowie Einflussfaktoren auf das Klima (Entstehung von Treibhausgasen, natürlicher und anthropogener Treibhauseffekt). Insgesamt konnten 40 Punkte erreicht werden. Weiterhin wurden die Dimensionen des Tests zum systemischen Denken untereinander als auch mit dem Gesamttest korreliert, um Argumente unterstützende Belege für die konvergente Validität zu sammeln. Zusätzlich wurden die Deutschnote der Lernenden und das Geschlecht erfasst, um Hinweise für die diskriminante Validität zu finden. Die

Deutschnote wurde herangezogen, da nicht Lesekompetenz, sondern systemisches Denken gemessen werden sollte, die Lesekompetenz aber durchaus einen ungewollten Einfluss haben könnte. Das Geschlecht wurde herangezogen, da in anderen Studien geschlechtsspezifische Unterschiede beobachtet wurden (Booth-Sweeney & Serman, 2000; Ossimitz, 2001). Jedoch sollte das Geschlecht aufgrund der geschlechtsneutralen Aufgaben keinen Einfluss auf die Testleistung haben. Um Hinweise auf die konvergente und diskriminante Validität zu erhalten, wurden zur Berechnung der Korrelationen mit den genannten Konstrukten die Scores vor und nach der Intervention verwendet. Potenzielle Konventionen für genauere Interpretationen können von Greiff, Stadler, Sonnleitner, Wolff und Martin (2015) und Fanta et al. (2017) entnommen werden. Hierbei sind

Korrelationen im Bereich  $r = .60$  bis  $r = .70$  als Hinweise auf konvergente Validität, Korrelationen von  $r = .25$  bis  $r = .45$  als Hinweise auf diskriminante Validität zu betrachten. Da allerdings Korrelationen von  $r = .45$  nach gängigen Konventionen (bspw. Schober, Boer & Schwarte, 2018) bereits als moderat eingestuft werden, wird angestrebt, noch geringere Korrelationen aufzuzeigen, um Hinweise auf die diskriminante Validität zu erhalten.

## 6.2 Ergebnisse

Die deskriptiven Itemkennwerte sind in Tabelle 3 dargestellt. Sie zeigt die Itemkennwerte für den Test zum systemischen Denken vor der Intervention ( $t1$ ) und nach der Intervention ( $t2$ ).

Tabelle 3

Hauptstudie: Itemkennwerte des Tests vor der Intervention ( $t1$ ) und nach der Intervention ( $t2$ )

Dimension /Item	Itemschwierigkeit	Varianz	Trennschärfe	Faktorladung
<b>Dimension 1 (Deklaratives/konzeptuelles systemisches Wissen)</b>				
Item 1 ( $t1$ )	.77	.07	.02	.67
Item 1 ( $t2$ )	.86	.05	.23	.32
Item 2 ( $t1$ )	.80	.20	.03	.01
Item 2 ( $t2$ )	.83	.14	.14	.09
Item 3 ( $t1$ )	.54	.15	.08	.31
Item 3 ( $t2$ )	.82	.09	.17	.64
<b>Dimension 2 (Systemmodellierungsfähigkeit)</b>				
Item 4 ( $t1$ )	.31	.21	.16	.14
Item 4 ( $t2$ )	.37	.23	.28	.24
Item 5 ( $t1$ )	.56	.06	.16	.59
Item 5 ( $t2$ )	.66	.06	.28	.68
<b>Dimension 3 (Lösen komplexer Probleme)</b>				
Erklärungen geben				
Item 6 ( $t1$ )	.19	.08	.30	.99
Item 6 ( $t2$ )	.42	.11	.43	1.21
Item 7 ( $t1$ )	.66	.39	.21	.64
Item 7 ( $t2$ )	.75	.07	.42	.51
Item 8 ( $t1$ )	.35	.15	.41	1.19
Item 8 ( $t2$ )	.55	.20	.45	1.12

Dimension /Item	Itemschwierigkeit	Varianz	Trennschärfe	Faktorladung
<b>Dimension 3 (Lösen komplexer Probleme)</b>				
Prognosen treffen				
Item 9 (t1)	.73	.16	.30	.63
Item 9 (t2)	.88	.08	.24	.44
Item 10 (t1)	.80	.09	.34	.51
Item 10 (t2)	.89	.06	.38	.41
Item 11 (t1)	.61	.13	.30	.52
Item 11 (t2)	.75	.11	.17	.42
Technologien entwickeln				
Item 12 (t1)	.19	.16	-	.16
Item 12 (t2)	.30	.21	-	.21

*Test vor der Intervention (t1).* Nach der Editierung der Items konnte das vorangegangene Problem der geringen Schwierigkeit der Items reduziert werden. Allerdings sind die Itemschwierigkeiten der ersten Dimension noch vergleichsweise niedrig. Die Itemschwierigkeiten der anderen Dimensionen liegen im akzeptablen, mittleren Bereich. Problematisch sind weiterhin die geringen Trennschärfen der ersten beiden Dimensionen. Die Trennschärfen der Items der dritten Dimension sind hierbei höher. Dies schlägt sich auch in den Reliabilitäten nieder: Dimension 1:  $\alpha = .08$ , Dimension 2:  $\alpha = .23$ , Dimension 3:  $\alpha = .63$  (Dimension 3: Erklärungen geben:  $\alpha = .44$ , Dimension 3: Prognosen treffen:  $\alpha = .49$ , gesamter Test:  $\alpha = .64$ ). Die Ergebnisse der KFA (t1) zeigen insgesamt eine gute Modellpassung, CFI = .92, RMSEA = .04. Allerdings sind die Faktorladungen in den Dimensionen 1 und 2 zu gering, was auf eine schwache Beziehung zwischen den Items 2, 3 und 4 und den jeweiligen Dimensionen hinweist. In Dimension 3 hingegen sind für die Items 6-11 durchgängig starke Faktorladungen über .50 zu finden, was darauf hinweist, dass diese die Dimension gut repräsentieren. Lediglich Item 12 weist eine deutlich zu geringe Faktorladung auf.

*Test nach der Intervention (t2).* Die Itemkennwerte haben sich nach der Intervention grundsätzlich verbessert. Die Itemschwierigkeiten sind geringer, was auf die Intervention und den damit verbundenen Wissenszuwachs zurückzuführen ist. Die Itemtrennschärfen sind grundsätzlich höher, wodurch auch die Reliabilitäten steigen. Jedoch bleiben die Reliabilitäten in den Kompetenzdimensionen 1 und 2 weiterhin auf einem niedrigen Niveau: Dimension 1:  $\alpha = .30$ , Dimension 2:  $\alpha = .37$ , Dimension 3:  $\alpha = .68$

(Dimension 3: Erklärungen geben:  $\alpha = .61$ ; Dimension 3: Prognosen treffen:  $\alpha = .41$ , gesamter Test:  $\alpha = .73$ ). Die Ergebnisse der KFA (t2) zeigen insgesamt eine gute Modellpassung, CFI = .98, RMSEA = .02. Jedoch zeigen sich, wie bereits zu t1, in den Dimensionen 1 und 2 teilweise geringe Faktorladungen (Items 1, 2 und 4), was auch hier auf eine schwache Beziehung zwischen Items und Dimensionen hinweist. In Dimension 3 liegen die Faktorladungen für die Items 6-11 im akzeptablen bis guten Bereich. Item 12 weist weiterhin eine deutlich zu geringe Faktorladung auf.

*Interventionseffekt.* t-Tests zeigten, dass sich die Schülerinnen und Schüler in jeder Dimension steigern konnten, Dimension 1:  $t(292) = 12.12, p < .001, d = 0.71$ , Dimension 2:  $t(292) = 5.73, p < .001, d = 0.45$ , Dimension 3:  $t(292) = 13.77, p < .001, d = 0.94$  (Dimension 3: Erklärungen geben:  $t(292) = 12.22, p < .001, d = 0.84$ , Dimension 3: Prognosen treffen:  $t(292) = 8.63, p < .001, d = 0.63$ , Dimension 3: Technologien entwickeln:  $t(292) = 3.51, p < .001, d = 0.32$ , gesamter Test:  $t(292) = 18.26, p < .001, d = 1.21$ ). Somit reagiert der überarbeitete Test, wie auch die ursprüngliche Testversion, sensitiv auf instruktionale Interventionen, wobei dieser Effekt angesichts der geringen Reliabilitätskennwerte in den Dimensionen 1 und 2 nach wie vor nur eingeschränkt interpretierbar ist.

*Validierung.* In Tabelle 4 sind die Korrelationskoeffizienten der Berechnungen der konvergenten Validität vor der Intervention (t1) und nach der Intervention (t2) dargestellt, in Tabelle 5 jene der diskriminanten Validität vor der Intervention (t1) und nach der Intervention (t2).

Tabelle 4

Konvergente Validität vor der Intervention (*t1*) und nach der Intervention (*t2*): Korrelationskoeffizienten

	Dimen- sion 1	Dimen- sion 2	Dimension 3				Gesam- ter Test	Klima- wissen
			ge- samt	Erklä- rungen	Prog- nosen	Techno- logie		
Dimension 1 ( <i>t1</i> )	-							
Dimension 1 ( <i>t2</i> )	-							
Dimension 2 ( <i>t1</i> )	.26*	-						
Dimension 2 ( <i>t2</i> )	.26*	-						
Dimension 3	gesamt ( <i>t1</i> )	.34*	.40*	-				
	gesamt ( <i>t2</i> )	.33*	.50*	-				
	Erklär- ungen ( <i>t1</i> )	.34*	.38*	.89*	-			
	Erklär- ungen ( <i>t2</i> )	.34*	.47*	.92*	-			
	Progno- sen ( <i>t1</i> )	.20*	.26*	.78*	.41*	-		
	Progno- sen ( <i>t2</i> )	.17*	.34*	.76*	.46*	-		
	Techno- logie ( <i>t1</i> )	.19*	.26*	.35*	.23*	.25*	-	
	Techno- logie ( <i>t2</i> )	.23*	.32*	.43*	.34*	.24*	-	
Gesamter Test ( <i>t1</i> )	.58*	.56*	.95*	.86*	.71*	.37*	-	
Gesamter Test ( <i>t2</i> )	.56*	.66*	.94*	.88*	.69*	.46*	-	
Klimawissen ( <i>t1</i> )	.36*	.39*	.57*	.56*	.36*	.29*	.62*	-
Klimawissen ( <i>t2</i> )	.33*	.43*	.65*	.32*	.42*	.32*	.67*	-

Anmerkung. Pearson Korrelationskoeffizienten sind dargestellt; \* $p < .05$ 

Tabelle 5

Diskriminante Validität vor der Intervention (*t1*) und nach der Intervention (*t2*): Korrelationskoeffizienten

	Deutschnote	Geschlecht
Dimension 1 ( <i>t1</i> )	.04	.03
Dimension 1 ( <i>t2</i> )	.05	-.01
Dimension 2 ( <i>t1</i> )	-.001	-.03
Dimension 2 ( <i>t2</i> )	.06	-.03

	Deutschnote	Geschlecht
Dimension 3	gesamt (t1)	-.03
	gesamt (t2)	<.001
	Erklärungen (t1)	-.02
	Erklärungen (t2)	.03
	Prognosen (t1)	-.01
	Prognosen (t2)	-.03
	Technologie (t1)	-.11
	Technologie (t2)	-.14*
Gesamter Test (t1)	-.01	.02
Gesamter Test (t2)	.03	-.06

*Anmerkung.* Pearson Korrelationskoeffizienten sind dargestellt für Korrelationen mit Deutschnote; Kendall-Tau-b Korrelationen sind dargestellt für Korrelationen mit dem Geschlecht; \* $p < .05$

Die konvergente Validität der Dimensionen konnte im Wesentlichen unterstützt werden. Es wurden durchweg positive signifikante Korrelationen gefunden. Dimension 1 korreliert mit vergleichsweise geringen Effektgrößen mit den anderen Facetten des systemischen Denkens und dem Fachwissen zum Klimawandel. Die Dimensionen 2 und 3 korrelieren sowohl stärker untereinander als auch mit dem Fachwissen. Grundsätzlich steigt die Korrelation nach der Intervention zwischen den Facetten mit dem Fachwissen von Dimension zu Dimension an, wenn man die Dimensionen global betrachtet. Dies trifft für die Korrelationen vor und nach der Intervention zu. Die diskriminante Validität wird bezüglich der Korrelationen nach der Intervention ebenfalls unterstützt. Es wurden nahezu ausschließlich Null-Korrelationen gefunden. Die einzige signifikante Korrelation wurde zwischen der Deutschnote und einer Teilkompetenz der dritten Dimension gefunden. Die signifikante Korrelation bedeutet hierbei, dass mit besserer Deutschnote die Scores in der Teilkompetenz höher ausfallen. Diese Korrelation ist allerdings sehr gering und weit unter den gängigen Grenzwerten. Die diskriminante Validität vor der Intervention wird ebenfalls unterstützt, da durchweg keine Korrelationen gefunden wurden. Abschließend wird die Konstruktvalidität dadurch unterstützt, dass alle drei Dimensionen Facetten desselben übergeordneten Konstrukts (systemisches Denken) darstellen. Die Korrelationen zwischen den Dimensionen und mit dem Fachwissen zeigen, dass sie unterschiedliche Aspekte dieses Konstrukts erfassen.

### 6.3 Diskussion

Auch nach Anhebung des Anforderungsniveaus in der finalen Testversion zeigen sich nach wie vor geringe bis mittlere Itemschwierigkeiten in den Dimensionen 1 und 2. Die Schwierigkeit steigt jedoch über die Dimensionen hinweg an, sodass sich das Lösen komplexer Probleme mit Hilfe von Systemmodellen in Dimension 3 als anspruchsvoll herausstellt. Die Kennwerte vor der Intervention waren auch in der zweiten Studie problematisch, vor allem die geringen Trennschärfen der unteren Dimensionen. Diese haben sich jedoch nach der Intervention sowie im Vergleich zur Pilotstudie verbessert. Das Messinstrument weist mit  $\alpha = .73$  eine gute interne Konsistenz auf (Cronbach, 1951). Die Reliabilitäten für die einzelnen Dimensionen fallen niedriger aus. Dies lässt sich zum einen durch die begrenzte Anzahl an Items erklären. Zum anderen wurden innerhalb der Dimensionen zwar ähnliche, aber doch unterschiedliche Konstrukte (Teilfähigkeiten) gemessen. Die Auswertung musste aufgrund der geringen Itemanzahl aber auf Ebene der übergeordneten Dimensionen und nicht der Teilfähigkeiten durchgeführt werden. Eine Erhöhung der Itemanzahl pro Teilfähigkeit würde zwar eine Auswertung auf dieser Ebene ermöglichen, wäre aufgrund der zunehmenden Testlänge für die Zielgruppe der Grundschule jedoch nicht mehr zumutbar, zumal dies ein Absinken der Konzentration und Motivation über die Zeit mit sich bringen würde. Abschließend lassen sich die Reliabilitäten der Kompetenzdimensionen 1 und 2 als unzureichend und die der Dimension 3 als ausreichend einstufen (Schmitt, 1996). Auch in Betracht der strukturellen Validität zeigen sich die

beiden unteren Kompetenzdimensionen als problematisch. Sowohl die geringen Faktorladungen einiger Items als auch die starke Varianz zwischen  $t1$  und  $t2$  weisen auf eine schwache und inkonsistente Beziehung zwischen den Items und Dimensionen hin. Dimension 3 zeigt hingegen durchgängig stärkere Faktorladungen und ist somit durch die Items besser repräsentiert. Hinweise auf konvergente Validität (Tab. 3) werden durch die steigenden Korrelationen zwischen den Dimensionen sowie deren Korrelationen mit dem Gesamtest und dem Fachwissen zum Klimawandel geliefert. Die Zunahmen in den Korrelationen bei höheren Kompetenzdimensionen deuten darauf hin, dass die höheren Dimensionen stärker mit dem übergeordneten Konstrukt verknüpft sind. Dieser Befund zeigte sich auch bei der Testentwicklung von Fanta et al. (2017) mit Lehramtsstudierenden, woraus sich ableiten lässt, dass auch das hier vorgestellte Messinstrument für die Primarstufe die verschiedenen Dimensionen des heuristischen Kompetenzstrukturmodells (Rieß et al., 2015) in Teilen abbildet und erfasst. Da die Teilfähigkeiten in Dimension 3 nochmals unterteilt wurden, zeigt sich bei der Entwicklung von Technologien eine vergleichsweise geringe Korrelation, was damit erklärt werden kann, dass es sich hierbei um ein exploratives Konstrukt handelt, welches mit nur einem Item erfasst wurde. Unterstützende Belege für die diskriminante Validität (Tab. 4) konnten insofern gesammelt werden, dass sich wünschenswerterweise keine Korrelation mit dem Geschlecht, wohl aber in geringem Maße mit der Deutschnote zeigt. Bei der Testkonstruktion wurde darauf geachtet, die Items möglichst sprachsensibel zu gestalten, gleichwohl das durch die Anforderungen des zugrundeliegenden Kompetenzstrukturmodells (Rieß et al., 2015) nur bedingt möglich ist. Es bedarf einer grundlegenden Lese- und Schreibkompetenz, um beispielsweise systemisch modellierbare Wirklichkeitsbereiche mit Hilfe eines Textes/Wortmodells verstehen (Dimension 2) oder mit Hilfe qualitativer/quantitativer Systemmodelle Erklärungen geben (Dimension 3) zu können.

## 7. Generelle Diskussion

Ziel dieser Studie waren die Entwicklung, empirische Prüfung und praktische Erprobung eines Messinstruments zur Erfassung systemischen Denkens von Viertklässlerinnen und Viertklässlern. Als Basis

wurde das heuristische Kompetenzstrukturmodell (Rieß et al., 2015) gewählt, da sich dieses bereits in höheren Jahrgangsstufen etabliert hat (Brockmüller, 2019; Fanta et al., 2017). Es sollte untersucht werden, ob dieses Modell auch auf den Primarbereich adaptiert werden kann, um eine Anschlussfähigkeit in der Förderung systemischen Denkens über die Schulformen hinweg zu erreichen. Die berichteten Ergebnisse zeigen, dass dies für die ausgewählten Teilfähigkeiten nur teilweise gelungen ist. Validitätsbetrachtungen haben gezeigt, dass der Test das Konstrukt misst, die angenommene Struktur (Dimensionalität) jedoch nur eingeschränkt abbildet. Diesbezüglich wurden Schwächen bei den beiden unteren Kompetenzdimensionen aufgrund geringer und inkonsistenter Faktorladungen festgestellt, während Dimension 3 gut durch die Items repräsentiert ist. Für die Problematik in den unteren Kompetenzdimensionen sehen wir zwei potenzielle Erklärungen, die weiterer Prüfung bedürfen. Zum einen könnte die geringe Itemanzahl die geringen Faktorladungen und die starke Varianz zwischen den Messzeitpunkten bedingen. Somit wären die Teilkonstrukte nicht ausreichend abgebildet. Vor diesem Hintergrund ist fraglich, ob systemisches Denken in der Primarstufe als umfassendes Konstrukt überhaupt adäquat messbar ist oder ob es sinnvoller wäre, einzelne Teilfähigkeiten zu adressieren. Die Ergebnisse zeigen aber, dass dies bei einer höheren Itemanzahl – wie in Dimension 3 – grundsätzlich möglich ist. Eine vollständige Erfassung würde jedoch erfordern, die Dimensionen 1 und 2 durch zusätzliche Items zu stärken. Allerdings könnte dies die Bearbeitungszeit verlängern und folglich die Konzentration und Motivation der Grundschülerinnen und Grundschüler beeinträchtigen. Zum anderen ist denkbar, dass die Dimensionalität des zugrundeliegenden Kompetenzmodells bei Schülerinnen und Schülern der Grundschule in dieser ausdifferenzierten Form noch nicht vorhanden ist. In der Pilotierungsstudie zeigten sich für den gesamten Test niedrige und auf Dimensionsebene inakzeptable Reliabilitäten, die jedoch durch die Überarbeitung und Eliminierung einzelner Items sowie die Ergänzung neuer Items für die Haupterhebung verbessert werden konnten. Dennoch waren auch noch in der Hauptstudie einzelne Items problematisch. Grund für die geringe Itemschwierigkeit könnte der fachliche Kontext des Klimawandels sein, da bei den Schülerinnen und Schülern durch die Omnipräsenz

des Themas im Alltag und in den Medien mehr Vorwissen vorhanden war als ursprünglich angenommen. Auch die angestrebte Sprachsensibilität des Tests könnte möglicherweise aufgrund zu einfacher Formulierungen der Items dazu geführt haben, dass sich diese als zu leicht erwiesen. Die größte Herausforderung ergab sich jedoch durch das Testformat. Zugunsten einer einfachen Handhabbarkeit und Auswertbarkeit wurden größtenteils Items im geschlossenen Format eingesetzt. Bei der Formulierung der Items stellte sich heraus, dass es schwierig ist, zu komplexen Systemen vereinfachte Aussagen im Format richtig oder falsch für die Zielgruppe der Grundschule zu treffen. Eine ähnliche Erfahrung machten auch Roczen, Fischer, Fögele, Hartig und Mehren (2024) bei der Entwicklung eines Tests zur Messung von nachhaltigkeitsbezogener Systemkompetenz bei Neuntklässlerinnen und Neuntklässlern. Abschließend lässt sich feststellen, dass wir uns bei der Entwicklung des Testinstruments zur Erfassung systemischen Denkens zum Klimawandel in der Primarstufe in einem Spannungsfeld bewegen. Um inhaltliche Anforderungen, Testumfang und Niveau angemessen zu berücksichtigen und zugleich psychometrische Gütekriterien zu erfüllen, sind entweder gezielte Kompromisse oder eine Fokussierung auf Teilfacetten systemischen Denkens erforderlich.

### **7.1 Stärken und Limitationen des Testinstruments**

Dieser Artikel verdeutlicht, wie herausfordernd die Testkonstruktion zur Messung eines mehrdimensionalen Konstrukts wie systemisches Denken angesichts der besonderen Anforderungen der Primarstufe ist. Bei der Datenauswertung hätten IRT-Analysen zugunsten der höheren Messgenauigkeit und detaillierteren Item-Analyse angewandt werden können. Jedoch wurden klassische Analyseverfahren unter dem Ziel der praktischen Erprobung des Testinstruments als angemessen und ausreichend betrachtet. Die Stärke dieser Studie liegt unter anderem darin, dass das Testinstrument im Rahmen einer Intervention erprobt wurde, was die praktische Tauglichkeit bestätigt. Die Sensitivität des Instruments erlaubt differenzierte Aussagen über den Erfolg der stattgefundenen Interventionen. So konnten auf Dimensionsebene unterschiedliche Wirkungen alternativer Systemmodelle zum Klimawandel ausgemacht werden. Allerdings ist die Instruktionssensitivität

angesichts geringer Reliabilitäten und der Probleme mit der strukturellen Validität in den unteren Kompetenzdimensionen nur eingeschränkt aussagekräftig. Vor diesem Hintergrund ist der Test in seiner aktuellen Form nicht geeignet, um die Dimensionalität systemischen Denkens ganzheitlich zu erfassen. Somit ist das Testinstrument in seiner jetzigen Fassung als vorläufig anzusehen und kann nur zur Messung der dritten Kompetenzdimension verlässlich eingesetzt werden. Zukünftige Forschung sollte neben der Optimierung und Weiterentwicklung der unteren Dimensionen auch die hier nicht beachteten Teilfähigkeiten des heuristischen Kompetenzstrukturmodells (Rieß et al., 2015) komplettieren – geachtet dessen, dass die vierte Kompetenzdimension für die Primarstufe wahrscheinlich zu komplex ist. Bei einer erneuten Evaluation sollten auch langfristige Effekte untersucht werden, wobei mehrere Messzeitpunkte mit ausreichend zeitlichem Abstand berücksichtigt werden, um die Auswirkungen auf die Itemschwierigkeiten zu erfassen. Das Testinstrument kann gerne durch Kontaktaufnahme mit dem Erstautor angefordert werden.

## Literatur

- Acher, A. (2017). *Welche Art von wissenschaftlichem Modellieren ist in der Grundschule angemessen?* Verfügbar unter: [https://opendata.uni-halle.de/bitstream/1981185920/94414/1/sachunterricht\\_volume\\_0\\_5642.pdf](https://opendata.uni-halle.de/bitstream/1981185920/94414/1/sachunterricht_volume_0_5642.pdf)
- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Arndt, H. (2017). *Systemisches Denken im Fachunterricht*. In H. Arndt (Hrsg.), *FAU Lehren und Lernen 2* (S. 9–25). Erlangen: FAU University Press.
- Ben-Zvi Assaraf, O. & Knippels, M. (2022). Lessons learned: Synthesizing Approaches That Foster Understanding of Complex Biological Phenomena. In O. Ben-Zvi Assaraf & M. Knippels (Hrsg.), *Fostering and Understanding of Complex Systems in Biology Education* (S. 249–278). Switzerland: Springer.
- Ben-Zvi Assaraf, O. & Orion, N. (2005). Development of System Thinking Skills in the Context of Earth System Education. *Journal of Research in Science Teaching*, 42(5), 518–560.
- Ben-Zvi Assaraf, O. & Orion, N. (2009). System thinking skills at the elementary school level. *Journal of Research in Science Teaching*, 47(5), 540–564.
- Bertschy, F. (2008). Vernetztes Denken in der Grundschule fördern. *Umweltpsychologie*, 12(2), 71–90.
- Bielik, T., Delen, I., Krell, M. & Ben-Zvi Assaraf, O. (2023). Characterising the literature on the teaching and learning of system thinking and complexity in STEM Education. *Journal for STEM Education Research*, 6(2), 199–231.
- Boersma, K., Waarlo A.J. & Klaassen, K. (2011). The feasibility of systems thinking in biology education. *Journal of Biological Education*, 45(4), 190–197.
- Bollmann-Zuberbühler, B., Frischknecht-Tobler, U., Kunz, P., Nagel, U. & Wilhelm Hamiti, S. (2010). *Systemdenken fördern. Systemtraining und Unterrichtsreihen zum vernetzten Denken. 1.–9. Schuljahr*. Bern: Schulverlag plus AG.
- Booth-Sweeney, L. & Serman, J.D. (2000). Bathtub Dynamics: Initial Results of a Systems Thinking Inventory. *System Dynamics Review*, 16(4), 249–286.
- Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin, Heidelberg: Springer.
- Bossel, H. (1992). *Simulation dynamischer Systeme: Grundwissen, Methoden, Programme*. Braunschweig: Vieweg.
- Bräutigam, J. (2014). *Systemisches Denken im Kontext einer Bildung für nachhaltige Entwicklung. Konstruktion und Validierung eines Messinstruments zur Evaluation einer Unterrichtseinheit* (Dissertation). Pädagogische Hochschule Freiburg. Verfügbar unter: <https://phfr.bsz-bw.de/frontdoor/deliver/index/docId/412/file/DissertationBraeutigamJulia2014.pdf>
- Brockmüller, S. (2019). *Erfassung und Entwicklung von Systemkompetenz - Empirische Befunde zu Kompetenzstruktur und Förderbarkeit durch den Einsatz analoger und digitaler Modelle im Kontext raumwirksamer Mensch-Umwelt-Beziehungen* (Dissertation). Pädagogische Hochschule Heidelberg. Verfügbar unter: [https://opus.ph-heidelberg.de/frontdoor/deliver/index/docId/340/file/Dissertation\\_Brockmueller\\_Systemkompetenz.pdf](https://opus.ph-heidelberg.de/frontdoor/deliver/index/docId/340/file/Dissertation_Brockmueller_Systemkompetenz.pdf)
- Brockmüller, S. (2023). Structure and measurement of system competence: Promoting systems thinking using analogue and digital models. In G.S. Carvalho, A.S. Afonso & Z. Anastácio (Hrsg.), *Fostering scientific citizenship in an uncertain world* (S. 79–94). Switzerland: Springer.
- Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230–258.
- Budak, U. & Ceyhan, G. (2023). Research trends on systems thinking approach in science education. *International Journal of Science Education*, 46(5), 485–502.
- Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.

- Carlson, K.D. & Herdman, A.O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, 15(1), 17–32.
- Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, 12(1), 25–44.
- Colliver, J.A., Conlee, M.J. & Verhulst, S.J. (2012). From test validity to construct validity... and back? *Medical Education*, 46(4), 366–371.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- DeVellis, R.F. & Thorpe, C.T. (2022). *Scale development: Theory and applications* (5. Aufl.). Thousand Oaks, California: SAGE Publications.
- Evagorou M., Korfiatis K., Nicolaou, C. & Constantinou, C. (2009). An investigation of the potential of interactive simulations for developing system thinking skills in elementary school: A case study with fifth-graders and sixth-graders. *International Journal of Science Education*, 31(5), 655–674.
- Fanta, D., Bräutigam, J., Greiff, S. & Rieß, W. (2017). Entwicklung und Validierung eines Messinstrumentes zur Erfassung von systemischem Denken bei Lehramtsstudierenden in ökologischen Kontexten. *Zeitschrift für Didaktik der Naturwissenschaften*, 23(1), 241–259.
- Fanta, D., Bräutigam, J. & Rieß, W. (2020). Fostering systems thinking in student teachers of biology and geography – an intervention study. *Journal of Biological Education*, 54(3), 226–244.
- Gesellschaft für Didaktik des Sachunterrichts (GDSU). (2013). *Perspektivrahmen Sachunterricht*. Bad Heilbrunn: Verlag Julius Klinkhardt.
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C. & Martin, R. (2015). Sometimes less is more. Comparing the validity of complex problem solving measures. *Intelligence*, 50, 100–113.
- Haider, M. (2019). *Modellkompetenz im naturwissenschaftlichen Sachunterricht. Eine empirische Studie zum Lernen mit Modellen und über Modelle in der Primarstufe*. Berlin: Logos Verlag.
- Hartig, J., Frey, A. & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 529–544). Heidelberg: Springer Berlin.
- Hmelo-Silver, C.E., Jordan, R., Eberbach, C. & Sinha, S. (2017). Systems learning with a conceptual representation: A quasi experimental study. *Instructional Science*, 45, 53–72.
- Hmelo-Silver, C.E. & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems and designs for learning. *Journal of the Learning Sciences*, 16(3), 307–331.
- Hokayem, H. & Gotwals, A.W. (2016). Early elementary students' understanding of complex ecosystems: A learning progression approach. *Journal of Research in Science Teaching*, 53(10), 1524–1545.
- Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- IBM Corporation. (2021). IBM SPSS Statistics for Windows (Version 28.0) [Computer software]. IBM Corp.
- Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Kinchin, I.M. (2020). A 'species identification' approach to concept mapping in the classroom. *Journal of Biological Education*, 54(1), 108–114.
- Klieme, E. & Maichle, U. (1991). *Erprobung eines Modellbildungssystems im Unterricht. Bericht über eine Pilotstudie zur Unterrichtsvaluation*. Bonn: Institut für Test- und Begabungsforschung.
- Koo, T.K. & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Krüger, D., Kauertz, A. & Upmeier zu Belzen, A. (2018). Modelle und das Modellieren in den Naturwissenschaften. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung*. (S. 141–158). Heidelberg: Springer Berlin.
- Lüschen, I. (2015). *Der Klimawandel in den Vorstellungen von Grundschulkindern. Wahrnehmung und Bewertung des globalen Umweltproblems*. Baltmannsweiler: Schneider-Verlag Hohengehren.
- Mambrey, S., Schreiber, N. & Schmiemann, P. (2020a). Young students reasoning about ecosystems: The role of systems thinking, knowledge, conceptions and representation. *Research in Science Education*, 52(1), 79–98.

- Mambrey, S., Timm, J., Landskron, J.J. & Schmiemann, P. (2020b). The impact of system specifics on systems thinking. *Journal of Research in Science Teaching*, 57(10), 1632–1651.
- Matthies, M. (2010). *Einführung in die Systemwissenschaft. Vorlesungsskript*. Osnabrück: Universität Osnabrück, Institut für Umweltsystemforschung.
- McDonald, R.P. (1999). *Test theory: A unified treatment*. New York: Lawrence Erlbaum Associates Publishers.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- Mehren, R., Rempfler, A., Ullrich-Riedhammer, E.-M., Buchholz, J. & Hartig, J. (2016). Systemkompetenz im Geographieunterricht. Ein theoretisch hergeleitetes und empirisch überprüftes Kompetenzstrukturmodell. *Zeitschrift für Didaktik der Naturwissenschaften*, 22, 147–163.
- Mischo, C. & Rieß, W. (2008). Förderung systemischen Denkens im Bereich von Ökologie und Nachhaltigkeit. *Unterrichtswissenschaft*, 36(4), 346–364.
- Moosbrugger, H. & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“) [Quality requirements for tests and questionnaires (“quality criteria”)]. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 13–38). Springer.
- Nationale Plattform Bildung für nachhaltige Entwicklung. (2017). Nationaler Aktionsplan. Bildung für nachhaltige Entwicklung. Der deutsche Beitrag zum UNESCO-Weltaktionsprogramm. Verfügbar unter: [https://www.bne-portal.de/bne/shareddocs/downloads/files/nationaler\\_aktionsplan\\_bildung\\_er\\_nachhaltige\\_entwicklung\\_neu.pdf?\\_\\_blob=publicationFile&v=3](https://www.bne-portal.de/bne/shareddocs/downloads/files/nationaler_aktionsplan_bildung_er_nachhaltige_entwicklung_neu.pdf?__blob=publicationFile&v=3)
- Nikendei, C., Bugaj, T.J., Nikendei, F., Kühl, S.J. & Kühl, M. (2020). Klimawandel: Ursachen, Folgen, Lösungsansätze und Implikationen für das Gesundheitswesen. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 156, 59–67.
- Nitz, S. & Fechner, S. (2018). Mentale Modelle. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 69–86). Heidelberg: Springer.
- Ossimitz, G. (2000). *Entwicklung systemischen Denkens. Theoretische Konzepte und empirische Untersuchungen*. München: Profil.
- Ossimitz, G. (2001). *Stock-Flow Thinking and Reading stock-flow-related Graphs: An Empirical Investigation in Dynamic Thinking Abilities*. Proceedings of the 2001 System Dynamics Conference, Atlanta, Georgia. Albany: System Dynamics Society.
- Rempfler, A. & Uphues, R. (2011a). Systemkompetenz und ihre Förderung im Geographieunterricht. *Geographie und Schule*, 33(189), 22–33.
- Rempfler, A. & Uphues, R. (2011b). Systemkompetenz im Geographieunterricht – Die Entwicklung eines Kompetenzmodells. In C. Meyer (Hrsg.), *Geographische Bildung: Kompetenzen in Forschung und Praxis. Gemeinsames Symposium des GEI und HGD* (S. 36–48). Braunschweig: Westermann.
- Rieß, W. & Mischo, C. (2009). Promoting systems thinking through biology lessons. *International Journal of Science Education*, 32(6), 705–725.
- Rieß, W. & Mischo, C. (2017). Das Modell problemorientierten Lehrens und Lernens (MopoLL) - Auf dem Weg zu einem evidenzbasierten Unterrichtsverfahren zur Förderung komplexer dynamischer Problemlösefähigkeiten in der Biologie. *Zeitschrift für Didaktik der Biologie*, 21, 1–20.
- Rieß, W., Schuler, S. & Hörsch, C. (2015). Wie lässt sich systemisches Denken vermitteln und fördern? Theoretische Grundlagen und praktische Umsetzung am Beispiel eines Seminars für Lehramtsstudierende. *Geographie aktuell und Schule*, 37(215), 16–29.
- Roczen, N., Fischer, F., Fögele, J., Hartig, J. & Mehren, R. (2024). Der „SysCo-ESD“ Test zur Messung von nachhaltigkeitsbezogener Systemkompetenz in Forschung und Praxis. In A. Rempfler, R. Grob, M. Landtwinning Blaser & U. Schönauer (Hrsg.), *Komplexität und Systemisches Denken im Geographieunterricht* (S. 86–101). Norderstedt: BoD – Books on Demand.
- Ropohl, M., Walpuski, M. & Sumfleth, E. (2015). Welches Aufgabenformat ist das richtige? – Empirischer Vergleich zweier Aufgabenformate zur standardbasierten Kompetenzmessung. *Zeitschrift für Didaktik der Naturwissenschaften*, 21, 1–15.

- Rosenkränzer, F., Stahl, E., Hörsch, C., Schuler, S. & Rieß, W. (2016). Das Fachdidaktische Wissen von Lehramtsstudierenden zur Förderung von systemischem Denken: Konzeptualisierung, Operationalisierung und Erhebungsmethode. *Zeitschrift für Didaktik der Naturwissenschaften*, 1(22), 109–121.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353.
- Schober, P., Boer, C. & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5), 1763–1768.
- Sommer, C. (2005). *Untersuchung der Systemkompetenz von Grundschulern im Bereich Biologie* (Dissertation). Christian-Albrechts-Universität zu Kiel. Verfügbar unter: [https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation\\_derivate\\_00001652/d1652.pdf](https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00001652/d1652.pdf)
- Stadler, M., Sailer, M. & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, 60, 100832.
- Streiling, S., Hörsch, C. & Rieß, W. (2021). Effects of Teacher Training in Systems Thinking on Biology Students—An Intervention Study. *Sustainability*, 13(14), 7631.
- Sweeney, L.B. & Sterman, J. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review*, 23(2–3), 285–311.
- Terzer, E. (2012). *Modellkompetenz im Kontext Biologieunterricht. Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice Items* (Dissertation). Humboldt-Universität zu Berlin. Verfügbar unter: <https://edoc.hu-berlin.de/server/api/core/bitstreams/3b7039d3-7f56-4596-b58d-c6f43f93df33/content>
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2017). *Education for Sustainable Development Goals. Learning Objectives*. Verfügbar unter: [https://www.unesco.de/sites/default/files/2018-08/unesco\\_education\\_for\\_sustainable\\_development\\_goals.pdf](https://www.unesco.de/sites/default/files/2018-08/unesco_education_for_sustainable_development_goals.pdf)
- Verhoeff, R.P., Knippels, M.-C.P.J., Gilissen, M.G.R. & Boersma, K.T. (2018). The theoretical nature of systems thinking. Perspectives on systems thinking in biology education. *Frontiers in Education*, 3, 518.
- Weise, G. (1975). *Psychologische Leistungstests*. Göttingen: Hogrefe.
- Westen, D. & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, 84(3), 608–618.
- Wiek, A. & Redman, A. (2022). What Do Key Competencies in Sustainability Offer and How to Use Them. In P. Vare, N. Lasselet, M. Rieckmann (Hrsg.), *Competences in Education for Sustainable Development* (Sustainable Development Goals Series, SDG: 4 Quality Education, S. 27–34). Cham: Springer.

## Kontakt

Sven Frey  
Pädagogische Hochschule Freiburg  
Institut für Biologie und ihre Didaktik  
Kunzenweg 21  
79117 Freiburg  
sven.frey@ph-freiburg.de

### Zitationshinweis:

Frey, S., Beege, M., Tramowsky, N. & Rieß, W. (2025). Entwicklung und Validierung eines Messinstruments zur Erfassung systemischen Denkens zum Klimawandel bei Grundschülerinnen und Grundschulern. *Zeitschrift für Didaktik der Biologie (ZDB) – Biologie Lehren und Lernen*, 30(1), 1-24. doi: 10.11576/zdb-7395

Veröffentlicht: 22.04.2025



Dieses Werk ist unter einer Creative Commons Lizenz vom Typ Namensnennung 4.0 International zugänglich (CC BY 4.0 de). Diese Lizenz gilt nur für das Originalmaterial. Alle gekennzeichneten Fremdinhalte (z.B. Abbildungen, Fotos, Tabellen, Zitate etc.) sind von der CC-Lizenz ausgenommen. Für deren Wiederverwendung ist es ggf. erforderlich, weitere Nutzungsgenehmigungen beim jeweiligen Rechteinhaber einzuholen. URL <https://creativecommons.org/licenses/by/4.0/>