

DIE 4K IM GEWAND EMPIRISCHER FORSCHUNG: EINE ANTWORT AUF ALBRECHT ET AL. (2025)

Benedikt Wisniewski
ehem. Universität Augsburg | benedikt.wisniewski@icloud.com

Tanja Fritz
Universität Augsburg | tanja.fritz@phil.uni-augsburg.de

Martin Daumiller
Universität Freiburg | martin.daumiller@psychologie.uni-freiburg.de

ABSTRACT

Die sogenannten „4K“ (Kommunikation, Kollaboration, kritisches Denken und Kreativität) spielen eine zentrale Rolle in aktuellen bildungspolitischen Diskursen und werden zunehmend empirisch modelliert. Der Beitrag von Albrecht et al. (2025) beansprucht, Struktur und Förderbarkeit dieser Kompetenzen im schulischen Kontext nachzuweisen. Die vorliegende Replik nimmt diesen Beitrag zum Anlass für eine systematische epistemische und methodische Prüfung. Im Zentrum steht die Frage, inwieweit die zugrunde gelegten Konstrukte theoretisch hinreichend bestimmt, empirisch trennbar und belastbar interpretierbar sind. Die Analyse zeigt, dass die berichteten Befunde die angenommene Struktur der 4K nur eingeschränkt stützen und dass die Prä-Post-Daten aufgrund des Designs keine starken Aussagen über Interventionseffekte erlauben. Am Beispiel von Albrecht et al. (2025) wird damit deutlich, welche besonderen Herausforderungen entstehen, wenn normativ etablierte Kompetenzmodelle in empirische Mess- und Interventionsforschung überführt werden. Wir plädieren entsprechend für eine klarere Trennung zwischen bildungspolitischer Zielsetzung, theoretischer Modellbildung und empirischer Prüfung.

SCHLAGWÖRTER

— 4K — 21ST CENTURY SKILLS — KOMPETENZMODELLIERUNG
— BILDUNGSFORSCHUNG — KONFIRMATORISCHE FAKTORENANALYSE —
REPLIKATIONSSTUDIE

ABSTRACT (ENGLISH)

THE 4CS IN THE GUISE OF EMPIRICAL RESEARCH: A REPLY TO ALBRECHT ET AL. (2025)

The so-called “4Cs” (communication, collaboration, critical thinking, and creativity) have become central concepts in contemporary educational policy and are increasingly operationalized in empirical research. The article by Albrecht et al. (2025) seeks to provide empirical evidence for the structure and malleability of these competencies in school contexts. This reply offers a systematic epistemological and methodological examination of this claim. It focuses on whether the underlying constructs are theoretically well defined, empirically distinguishable, and amenable to robust interpretation. The analysis shows that the reported findings provide only limited support for the proposed four-factor structure and that the pre–post design does not allow strong conclusions regarding intervention effects. Using the study by Albrecht et al. (2025) as an example, the paper illustrates the challenges that arise when normatively established competency frameworks are translated into empirical measurement and intervention research. It argues for a clearer distinction between educational policy objectives, theoretical model development, and empirical testing.

KEYWORDS

— 4Cs — 21ST CENTURY SKILLS — COMPETENCY MODELLING
— EDUCATIONAL RESEARCH — CONFIRMATORY FACTOR ANALYSIS —
REPLICATION STUDY

1 — EINLEITUNG

Konzepte wie die sogenannten „4K“ – Kommunikation, Kollaboration, kritisches Denken und Kreativität – haben in den vergangenen zwei Jahrzehnten eine zentrale Stellung in bildungspolitischen Diskursen eingenommen. Sie fungieren zunehmend als Leitkategorien für die Ausgestaltung von Curricula, Unterrichtsformaten und Prüfungsformen und werden häufig unter dem Sammelbegriff der „21st Century Skills“ verhandelt. Ihre Verbreitung ist eng mit internationalen Organisationen und Netzwerken wie der OECD sowie Initiativen wie der Partnership for 21st Century Learning (P21) verbunden, die entsprechende Kompetenzrahmenwerke entwickelt und in nationale Bildungssysteme eingespeist haben. Diese Rahmenwerke sind primär programmatischer Natur: Sie formulieren Zielvorstellungen für Bildung unter Bedingungen gesellschaftlichen Wandels, insbesondere im Kontext von Digitalisierung, Globalisierung und ökonomischer Transformation (OECD 2018; P21 2009).

Damit ist jedoch zunächst nur der normative Status dieser Konzepte beschrieben. Kompetenzrahmen wie die 4K legen fest, welche Fähigkeiten als wünschenswert gelten, nicht aber, ob und in welcher Weise diese Fähigkeiten als eigenständige, empirisch trennbare und messbare Konstrukte existieren. Die Überführung solcher Zielkategorien in wissenschaftliche Forschung ist daher eine eigenständige erkenntnistheoretische Herausforderung. In der empirischen Bildungsforschung wird diese Herausforderung typischerweise durch Operationalisierung, Messmodellierung und Interventionsevaluation adressiert. Dabei wird implizit oder explizit angenommen, dass sich die postulierten Kompetenzen reliabel erfassen, voneinander abgrenzen und gezielt fördern lassen.

Der Beitrag von Albrecht et al. (2025) bietet einen geeigneten Ausgangspunkt, um dieser Frage nachzugehen. Er ist im Kontext des Modellprojekts *LuPe²* („Zeitgemäße Lern- und Prüfungskultur entwickeln und erforschen“) angesiedelt, das die Förderung der 4K in schulischen Lehr-Lern- und Prüfungskontexten zum Ziel hat und zugleich empirisch begleitet wird.

Die Autorinnen und Autoren beanspruchen, empirisch fundierte Aussagen zur Struktur und Förderung der 4K im schulischen Kontext zu liefern. Dabei greifen sie auf ein Design zurück, das sowohl Interventionselemente als auch quantitative Modellierung umfasst, insbesondere in Form konfirmatorischer Faktorenanalysen. Der Beitrag steht damit exemplarisch für einen Forschungszugang, der darauf abzielt, etablierte Kompetenzrahmen empirisch zu operationalisieren und ihre pädagogische Relevanz zu untersuchen.

Ziel der vorliegenden Replik ist eine epistemische und methodologische Klärung. Im Zentrum steht nicht die politische oder pädagogische Bewertung der 4K, sondern die Frage, inwiefern der Beitrag von Albrecht et al. die Voraussetzungen theoriegeleiteter empirischer Forschung einlöst, die er adressiert. Dabei geht es weniger um einzelne Ergebnisse oder Detailentscheidungen als um die übergeordnete Forschungslogik.

Die Kritik soll zur Klärung beitragen, wo die Grenze zwischen Forschung, Reformrhetorik und Implementationsbegleitung verlaufen. Die zentrale These lautet, dass sich im Beitrag exemplarisch Spannungen zwischen normativ etablierten Kompetenzmo-

dellen und deren empirischer Prüfung zeigen. Insbesondere prüfen wir, inwieweit das zugrundeliegende Konzept tatsächlich zur Disposition steht oder in wesentlichen Teilen vorausgesetzt wird. Vor diesem Hintergrund analysieren wir, insbesondere durch eine Sekundäranalyse der bereitgestellten Daten, welche Rolle empirische Verfahren in der Studie einnehmen und in welchem Maße sie zur Prüfung, Differenzierung oder Absicherung der theoretischen Annahmen beitragen. Diese Überlegungen werden im Folgenden auf theoretischer, methodischer und empirischer Ebene entfaltet.

2 — DIE 4K ALS BILDUNGSPOLITISCH GENERIERTER KOMPETENZRAHMEN

Die „4K“ – Kommunikation, Kollaboration, kritisches Denken und Kreativität – entstammen keiner einheitlichen wissenschaftlichen Theorie, sondern bildungspolitischen Diskursen. Ihre Ausarbeitung erfolgt maßgeblich in internationalen Kompetenzrahmen wie jenen der OECD und der Partnership for 21st Century Learning (P21), die seit den frühen 2000er-Jahren Zielkompetenzen für Bildungssysteme definieren (OECD 2018; P21 2009).

Diese Rahmenwerke haben primär eine orientierende und steuernde Funktion: Sie bündeln bildungspolitische Zielvorstellungen, machen sie international vergleichbar und regen Reformen an. Entsprechend sind sie normativ angelegt, indem sie festlegen, welche Kompetenzen als wünschenswert gelten und welche Rolle Bildung im gesellschaftlichen Wandel spielen soll. Ihre Geltung beruht weniger auf empirischer Prüfung als auf politischer Aushandlung und institutioneller Autorität (Grek 2009; Sellar / Lingard 2013).

Daraus folgt nicht, dass die 4K wissenschaftlich ungeeignet oder pädagogisch irrelevant wären. Es bedeutet jedoch, dass es sich dabei um normative Zielkategorien handelt. Sie liefern keine erklärende Theorie darüber, wie diese Kompetenzen strukturiert sind, wie sie sich voneinander abgrenzen lassen und unter welchen Bedingungen sie sich entwickeln. Wissenschaftstheoretisch handelt es sich nicht um erklärende Theorien, sondern um ein normatives Modell, das empirischer Forschung bedarf.

Ihre Übernahme in die empirische Bildungsforschung ist deshalb anspruchsvoll. Sie setzt voraus, dass die Kategorien als latente Variablen rekonstruierbar, voneinander abgrenzbar und messbar sind – Annahmen, die nicht vorausgesetzt, sondern theoretisch begründet und empirisch geprüft werden müssen. Aus psychometrischer Perspektive erfordert dies insbesondere klar definierte und unterscheidbare Konstrukte (Borsboom 2005; Brown 2015).

Zugleich erfüllen Kompetenzkonzepte in der Bildungsforschung häufig eine doppelte Funktion: Sie strukturieren empirische Forschung und sind zugleich Ausdruck normativer Setzungen (Biesta 2007; Weinert 2001). Gerade deshalb ist die Unterscheidung zwischen programmatischer Verwendung und wissenschaftlicher Modellierung zentral. Allenfalls besteht die Gefahr, dass bildungspolitisch plausible Kategorien vorschnell als empirische Konstrukte behandelt werden.

Für die Analyse empirischer Studien zu den 4K ist diese Differenz entscheidend. Sie betrifft die Frage, ob Untersuchungen tatsächlich zur Klärung von Struktur, Messbarkeit und Förderbarkeit beitragen oder sie primär die Plausibilität bestehender Kompetenzrahmen empirisch stabilisieren. Die folgende Analyse greift dies auf und zeigt am Beispiel von Albrecht et al. (2025), wie sich die Spannung zwischen normativer Setzung und empirischem Anspruch manifestiert.

3 — WISSENSCHAFTLICHE MODELLIERUNG DER 4K: ANSPRUCH UND VORAUSSETZUNGEN

Die Überführung eines normativ gesetzten Kompetenzmodells in empirische Forschung erfordert eine Reihe theoretischer und methodischer Klärungen. Die darin enthaltenen Kategorien müssen so präzisiert werden, dass eine empirische Prüfung überhaupt ermöglicht wird. Dies betrifft insbesondere drei miteinander verbundene Anforderungen: die theoretische Bestimmung der Konstrukte, ihre Operationalisierung in Form messbarer Indikatoren sowie die empirische Prüfung ihrer Struktur.

Erstens setzt die empirische Untersuchung voraus, dass die postulierten Kompetenzen hinreichend klar definiert sind. Latente Konstrukte können nur dann sinnvoll modelliert werden, wenn theoretisch bestimmt ist, wodurch sie sich von anderen Konstrukten unterscheiden. In der psychometrischen Literatur wird diese Anforderung als Voraussetzung von Konstruktvalidität und diskriminanter Validität beschrieben (AERA / APA / NCME 2014; Borsboom 2005; Brown 2015). Für die 4K bedeutet dies, dass begründet werden muss, inwiefern Kommunikation, Kollaboration, kritisches Denken und Kreativität (und entsprechende Subdimensionen) eigenständige, voneinander abgrenzbare Dimensionen darstellen und nicht lediglich unterschiedliche Bezeichnungen für übergeordnete Fähigkeiten, allgemeine Selbstwirksamkeit, sprachliche Artikulationsfähigkeit, sozial erwünschte Dispositionen oder kontextabhängige Verhaltensweisen sind.

Zweitens erfordert die empirische Modellierung eine nachvollziehbare Operationalisierung. Die Entwicklung von Messinstrumenten setzt voraus, dass theoretische Annahmen darüber vorliegen, wie sich die jeweiligen Kompetenzen in beobachtbaren Indikatoren ausdrücken. Itemformulierungen, Skalenkonstruktion und Messmodelle müssen dabei aus der Theorie abgeleitet werden und nicht umgekehrt (DeVellis 2017). Dabei ist auch zu klären, ob die gewählten Indikatoren als reflektive Manifestationen eines gemeinsamen latenten Faktors verstanden werden oder ob sie eher unterschiedliche, formative Bestandteile eines breiteren Kompetenzbereichs darstellen. Werden Items ohne klare theoretische Fundierung entwickelt oder direkt aus bestehenden Frameworks übernommen, besteht die Gefahr, dass die Messung nicht das intendierte Konstrukt erfasst, sondern lediglich sprachliche oder normative Zuschreibungen reproduziert.

Drittens schließlich muss die angenommene Struktur der Konstrukte empirisch überprüft werden. Verfahren wie die konfirmatorische Faktorenanalyse dienen genau diesem Zweck: Sie ermöglichen es, vorab formulierte Modellannahmen an Daten zu testen und gegebenenfalls zu verwerfen (Brown 2015). Entscheidend ist dabei der

hypothetische Charakter des Vorgehens. Konfirmatorische Analysen sind nur dann sinnvoll, wenn sie die Möglichkeit einschließen, dass ein Modell nicht durch die Daten gestützt wird. Die empirische Nicht-Bestätigung eines Modells ist in diesem Sinne kein methodisches Problem, sondern ein zentraler Bestandteil wissenschaftlicher Erkenntnis (Popper 2002).

Aus diesen Anforderungen ergibt sich ein klarer Prüfmaßstab für empirische Studien zu den 4K. Solche Studien müssten zeigen, dass (a) die postulierten Kompetenzen theoretisch hinreichend bestimmt sind, (b) ihre Operationalisierung auf dieser theoretischen Grundlage erfolgt und (c) die empirischen Analysen tatsächlich geeignet sind, die zugrunde liegenden Modellannahmen zu testen – einschließlich der Möglichkeit ihrer Widerlegung.

Gleichzeitig ist zu berücksichtigen, dass die Ausgangskonzepte aus bildungspolitisch-normativen Kontexten stammen. Dies erhöht die Anforderungen an ihre wissenschaftliche Modellierung zusätzlich. Während in theoriegeleiteter Forschung die Konstrukte aus bestehenden theoretischen Ansätzen abgeleitet werden, erfolgt hier der umgekehrte Weg: Ein bereits etabliertes, politisch verbreitetes Konzept soll nachträglich empirisch fundiert werden. Dadurch stellt sich besonders deutlich die Frage, ob empirische Verfahren ergebnisoffen zur Prüfung des Modells eingesetzt werden oder ob sie vor allem dessen Plausibilität im Forschungsfeld absichern.

Vor diesem Hintergrund ist es für die Bewertung entsprechender Studien entscheidend zu klären, wie empirische Verfahren hierzu tatsächlich eingesetzt werden. Sie können zur Klärung, Differenzierung und Revision eines Kompetenzmodells beitragen; sie können aber auch programmatische Kategorien in empirischer Form fort-schreiben. Diese Unterscheidung reicht über den Beitrag von Albrecht et al. (2025) hinaus: Sie betrifft grundsätzlich die Frage, wie bildungspolitisch etablierte Kompetenzrahmen wissenschaftlich untersucht werden können, ohne ihre normative Plausibilität vorschnell als empirische Gültigkeit zu behandeln. Diese Unterscheidung bildet den Ausgangspunkt für die folgende Analyse.

4 — ANALYSE DES BEITRAGS VON ALBRECHT ET AL. (2025)

4.1 FORSCHUNGSLOGIK UND DESIGN

Der Beitrag von Albrecht et al. (2025) beansprucht, empirisch fundierte Aussagen zur Struktur und Förderung der sogenannten 4K im schulischen Kontext zu liefern. Methodisch kombiniert die Studie qualitative Begleitforschung mit quantitativen Analysen, insbesondere in Form konfirmatorischer Faktorenanalysen sowie Prä-Post-Vergleichen im Rahmen eines Interventionsprojekts. Damit wird der Anspruch formuliert, sowohl die empirische Struktur des Konstrukts als auch dessen pädagogische Wirksamkeit zu untersuchen. Der Beitrag formuliert dabei explizit auch einen Validierungsanspruch. So halten Albrecht et al. (2025) fest, mit dem entwickelten Fragebogen liege „ein Messinstrument vor, welches für das Fach Deutsch und zum Teil auch für Mathematik, Englisch und Kunst grundlegend validiert wurde“.

4.1.1 VORSTRUKTURIERUNG DES ERKENNTNISINTERESSES

Bereits auf der Ebene der Problemformulierung wird deutlich, dass die Relevanz der 4K nicht als zu prüfende Annahme behandelt wird, sondern als Ausgangspunkt der Untersuchung dient. Die Förderung dieser Kompetenzen erscheint im Beitrag nicht als zu prüfende Möglichkeit, sondern als pädagogisch gebotenes Ziel. Entsprechend sind zentrale Begriffe wie „zeitgemäße Lernkultur“, „notwendige Kompetenzen“ oder „zukunftsrelevante Fähigkeiten“ nicht analytisch bestimmt, sondern fungieren als normative Prämissen.

Damit verschiebt sich die Funktion empirischer Forschung. An die Stelle einer Hypothesenprüfung tritt eine Form der Bestätigungsforschung, in der empirische Verfahren primär dazu dienen, die Plausibilität eines vorab gesetzten Reformanliegens zu stützen. In der Methodologie empirischer Sozialforschung wird eine solche Konstellation als problematisch beschrieben, da sie die Möglichkeit systematischer Falsifikation einschränkt (Shadish et al. 2002). Je stärker ein Modell bereits als pädagogischer Orientierungsrahmen fungiert, desto wichtiger wird daher die Frage, ob und wie alternative Deutungen der Befunde systematisch berücksichtigt werden.

4.1.2 DESIGN ALS IMPLEMENTATIONSFORSCHUNG

Diese Vorstrukturierung setzt sich im Forschungsdesign fort. Die Studie ist in ein Projekt (LuPe²) eingebettet, das zugleich Fortbildungsmaßnahme, Schulentwicklungsinitiative und wissenschaftliche Begleitforschung ist. Lehrkräfte werden im Rahmen des Projekts gezielt für die Bedeutung der 4K sensibilisiert, entwickeln entsprechende Unterrichtsformate und reflektieren ihre Erfahrungen im Anschluss. Schülerinnen und Schüler werden parallel dazu zu genau jenen Kompetenzdimensionen befragt, die zuvor als Ziel der Intervention definiert wurden.

Ein solches Design entspricht der Logik von Implementations- oder Evaluationsforschung, weil es die Einführung und praktische Ausgestaltung eines pädagogischen Konzepts begleitet (Rossi et al. 2004; Patton 2015). Für die Prüfung theoretischer Strukturannahmen und kausaler Effekte ist es jedoch nur eingeschränkt brauchbar. Insbesondere fehlt es an Vergleichsbedingungen, mit denen alternative Erklärungen für beobachtete Veränderungen systematisch ausgeschlossen werden könnten.

4.1.3 DATENGRUNDLAGE UND MESSLOGIK

Die quantitative Erhebung basiert auf standardisierten Selbsteinschätzungen von Schülerinnen und Schülern, die zu zwei Messzeitpunkten im Rahmen eines Prä-Post-Designs erhoben wurden. Erfasst wurden Einschätzungen zu den vier Kompetenzbereichen Kommunikation, Kollaboration, kritisches Denken und Kreativität sowie zugehörige Subdimensionen (u. a. Wissen/Fähigkeiten, Einstellungen und wahrgenommene Gelegenheiten).

Die Datenerhebung ist damit auf subjektive Einschätzungen der Lernenden ausgerichtet und zielt darauf ab, Veränderungen in diesen Einschätzungen im zeitlichen

Verlauf abzubilden. Die Auswertung erfolgt über latente Variablenmodelle sowie Mittelwertvergleiche zwischen den Messzeitpunkten.

4.2 METHODISCHE PROBLEME

Bevor die methodischen Einschränkungen im Einzelnen diskutiert werden, ist hervorzuheben, dass der Beitrag in mehrfacher Hinsicht Transparenz herstellt. Positiv zu würdigen ist insbesondere, dass Albrecht et al. (2025) ergänzende methodische Informationen, detaillierte Ergebnisse, Daten und Auswertungsskripte öffentlich bereitstellen. Erst dadurch wird eine differenzierte Prüfung der Analysen überhaupt möglich. Die folgende Kritik nutzt diese Transparenz, um die Reichweite der empirischen Befunde genauer zu bestimmen.

4.2.1 KONFIRMATORISCHE FAKTORENANALYSE: VORAUSSETZUNGEN UND INTERPRETATIONSGRENZEN

Der zentrale quantitative Anspruch des Beitrags von Albrecht et al. (2025) liegt in der empirischen Prüfung der postulierten Struktur der 4K mittels konfirmatorischer Faktorenanalyse (CFA). Dieses Verfahren ist grundsätzlich geeignet, theoretisch spezifizierte Modelle latenter Variablen zu testen – sofern die Modellannahmen klar definiert sind, die Datengrundlage den Anforderungen genügt und die Ergebnisse methodisch korrekt interpretiert werden (Brown 2015; Kline 2016).

Im vorliegenden Fall sind diese Voraussetzungen in mehrfacher Hinsicht nicht erfüllt, was die Aussagekraft der Ergebnisse insgesamt infrage stellt. Dies betrifft die Reproduzierbarkeit der Datenaufbereitung, die Interpretation statistischer Modellhinweise, den Umgang mit nachträglichen Modellanpassungen sowie die empirische Trennbarkeit der postulierten Dimensionen.

4.2.2 FEHLENDE REPRODUZIERBARKEIT DER DATENGRUNDLAGE

Bereits auf der Ebene der Datenaufbereitung ergeben sich Fragen der Nachvollziehbarkeit: Im Analyseprozess treten beim Zusammenführen der Längsschnittdaten many-to-many-Relationen auf, die durch manuelle Eingriffe außerhalb des dokumentierten Skripts „korrigiert“ werden. Damit ist nicht mehr nachvollziehbar, welche Datenbasis den Analysen tatsächlich zugrunde liegt.

Reproduzierbarkeit ist ein zentrales Gütekriterium quantitativer Forschung. Sämtliche Datenbereinigungs- und Transformationsschritte müssen vollständig dokumentiert und idealerweise skriptbasiert nachvollziehbar sein (Peng 2011). Manuelle, nicht dokumentierte Eingriffe unterlaufen dieses Prinzip. Die resultierenden Analysen sind damit nicht überprüfbar und verlieren ihren wissenschaftlichen Status als replizierbare Befunde.

4.2.3 MISSINTERPRETATION ZENTRALER MODELLWARNUNGEN

Ein weiteres Problem betrifft den Umgang mit statistischen Warnmeldungen. Im Analyseoutput wird explizit darauf hingewiesen, dass die Kovarianzmatrix der latenten Variablen nicht positiv definit ist. In der Strukturgleichungsmodellierung gilt dies als klarer Hinweis auf Modell-Misspezifikation, extreme Multikollinearität oder fehlende Trennbarkeit der Faktoren (Kline 2016).

Diese Warnung wird im Skript jedoch als unproblematisch interpretiert, da „alle Werte positiv“ seien. Diese Interpretation ist fachlich falsch. Positive Definitheit ist keine Eigenschaft einzelner Varianzen oder Kovarianzen, sondern der gesamten Matrix. Sie hängt von der Eigenwertstruktur ab und kann auch dann verletzt sein, wenn alle Einträge positiv sind.

Inhaltlich kann eine nicht positiv definite Kovarianzmatrix darauf hindeuten, dass die angenommenen latenten Faktoren empirisch nicht unterscheidbar sind oder in einer linearen Abhängigkeit stehen. Genau dies ist hier der Fall: Die latenten Korrelationen zwischen zentralen Dimensionen (insbesondere Wissen und Fähigkeiten) liegen im Bereich von $r > .90$ und sprechen dafür, dass die betreffenden Faktoren empirisch kaum unterscheidbar sind. Solche Werte gelten als klarer Hinweis auf fehlende diskriminante Validität und Redundanz der Konstrukte.

Statt diesen Befund als Widerlegung der Modellannahme zu interpretieren, wird er ignoriert. Damit wird ein zentraler diagnostischer Hinweis, der die Tragfähigkeit des gesamten Messmodells betrifft, methodisch nicht verarbeitet.

4.2.4 KONFIRMATORISCHE ANALYSE ALS DATENGETRIEBENE MODELLRETTUNG

Ein weiteres Problem ergibt sich aus dem Umfang nachträglicher Modellanpassung durch die systematische Nutzung von Modifikationsindizes. Im Analyseprozess werden wiederholt zusätzliche Parameter freigesetzt, insbesondere residuale Kovarianzen zwischen Items sowie Intercepts im Rahmen partieller Messinvarianz.

Solche Eingriffe sind nicht grundsätzlich unzulässig, markieren jedoch einen klaren Übergang von konfirmatorischer zu explorativer Modellierung. Modifikationsindizes sind hypothesengenerierend, nicht hypothesenprüfend (MacCallum et al. 1992). Ihre extensive Nutzung führt dazu, dass das Modell schrittweise an die Daten angepasst wird, anstatt die Daten zur Prüfung des Modells zu verwenden. Die resultierenden Modelle sind dann eher als empirisch angepasste Arbeitsmodelle zu verstehen und weniger als strenge konfirmatorische Prüfung eines vorab spezifizierten Kompetenzmodells.

4.2.5 HEYWOOD-NAHE LÖSUNGEN UND MODELLINSTABILITÄT

Zusätzlich treten standardisierte Faktorladungen an der theoretischen Obergrenze (≈ 1.00) auf. Solche Werte gelten in der SEM-Literatur als Hinweis auf problemati-

sche Modellkonstellationen, etwa Multikollinearität, Überfaktorisierung oder Identifikationsprobleme (Kline 2016).

In Kombination mit den hohen latenten Korrelationen und Hinweisen auf eine nicht positiv definiter Kovarianzmatrix ergibt sich ein konsistentes Bild eingeschränkter Modellstabilität. Die angenommenen Faktoren sind empirisch nicht trennbar, und die Parameterschätzungen bewegen sich an den Grenzen des zulässigen Bereichs.

Diese Befunde legen nahe, dass die angenommenen Faktoren empirisch nicht trennbar sind. Sie sprechen gegen eine Interpretation der postulierten Subdimensionen als klar unterscheidbare latente Konstrukte. Für die weitere Instrumentenentwicklung wäre daher zu prüfen, ob alternative Messmodelle – etwa übergeordnete Faktoren, bifaktorielle Modelle, stärker bereichsspezifische Modelle oder eine Unterscheidung nach Itemfunktionen – besser zur Datenstruktur passen.

4.2.6 POST-HOC-ZUSAMMENLEGUNG VON FAKTOREN

Besonders aufschlussreich ist der Umgang mit der faktoriellen Nicht-Trennbarkeit der Subskalen „Wissen“ und „Fähigkeiten/Fertigkeiten“. Die empirischen Analysen zeigen, dass diese in keinem der 4K-Bereiche unterscheidbar sind. Dieser Befund widerspricht direkt der theoretischen Modellannahme.

Die nachträgliche Zusammenlegung beider Subskalen ist als pragmatische Reaktion auf die empirische Nicht-Trennbarkeit nachvollziehbar. Sie sollte jedoch nicht als Bestätigung des ursprünglichen Modells interpretiert werden. Vielmehr zeigt sie, dass die theoretisch angenommene Differenzierung zwischen Wissen und Fähigkeiten/Fertigkeiten durch die vorliegenden Daten nicht gestützt wird. Die neu gebildete Skala „Wissen/Fähigkeiten/Fertigkeiten“ bedarf daher einer eigenständigen theoretischen Begründung: Es müsste geklärt werden, ob sie ein übergeordnetes Kompetenzselbstkonzept, eine allgemeine Fähigkeitswahrnehmung oder eine Mischdimension aus Wissen, Können und Selbstbeschreibung erfasst.

4.2.7 STICHPROBENGROSSE UND MODELLKOMPLEXITÄT

Schließlich erscheint auch die Stichprobengröße im Verhältnis zur Modellkomplexität problematisch. Nach dem veröffentlichten Analyseskript umfasst das längsschnittliche CFA-Modell einschließlich der Messinvarianzprüfung 118 frei zu schätzende Parameter, denen lediglich 90 Beobachtungen gegenüberstehen. Die erforderliche Stichprobengröße hängt zwar von der Modellkomplexität, den Faktorladungen und der Datenstruktur ab (Brown 2015; Kline 2016), gleichwohl liegt das Verhältnis von Parametern zu Beobachtungen deutlich unter klassischen Orientierungswerten. So empfehlen Bentler und Chou (1987) für Strukturgleichungsmodelle mindestens fünf, besser zehn Beobachtungen pro frei zu schätzendem Parameter. Bezogen auf das vorliegende Modell entspräche dies einer Stichprobengröße von etwa 590 bis 1180 Fällen. Auch wenn solche Faustregeln nicht schematisch anzuwenden sind, verdeutlicht der Vergleich, dass die Schätzungen auf einer außerordentlich schmalen empirischen Basis beruhen.

rischen Grundlage beruhen. Die Stabilität der Parameterschätzungen, Standardfehler und Modellfit-Indizes ist daher zusätzlich mit Vorsicht zu beurteilen.

4.2.8 VERÄNDERUNGSMESSUNGEN MIT BEGRENZTER AUSSAGEKRAFT DESIGNPROBLEME UND FEHLINTERPRETATIONEN

Neben der faktorenanalytischen Modellierung bildet die Analyse von Prä–Post-Veränderungen einen zweiten zentralen Pfeiler des empirischen Anspruchs von Albrecht et al. (2025). Ziel ist es, Aussagen über die Wirkung der durchgeführten Intervention auf die Ausprägung der 4K zu treffen. Die methodische Grundlage hierfür ist ein einfaches Prä–Post-Design, bei dem Selbsteinschätzungen von Schülerinnen und Schülern zu zwei Messzeitpunkten verglichen werden.

Auch dieser Teil der Analyse weist grundlegende methodische Probleme auf, die sowohl die interne Validität der Befunde als auch ihre Interpretation betreffen.

Auf Grund fehlender Vergleichsbedingungen ist die interne Validität beeinträchtigt: Das zentrale Problem des Designs besteht im Fehlen einer Kontroll- oder Vergleichsgruppe. Veränderungen zwischen Messzeitpunkt 1 und Messzeitpunkt 2 werden implizit als Effekte der Intervention interpretiert, obwohl alternative Erklärungen nicht ausgeschlossen werden können.

In der Methodologie experimenteller und quasi-experimenteller Forschung ist gut dokumentiert, dass Prä–Post-Designs ohne Kontrollgruppe eine Vielzahl systematischer Verzerrungen aufweisen (Campbell / Stanley 1963; Shadish et al. 2002). Dazu zählen insbesondere:

- Reifungseffekte: Veränderungen, die unabhängig von der Intervention durch Entwicklung oder Erfahrung entstehen
- Testeffekte: Veränderungen, die durch wiederholte Messung selbst verursacht werden
- Regression zur Mitte: statistische Artefakte bei extremen Ausgangswerten
- Kontexteffekte: Einflüsse durch schulische oder außerschulische Ereignisse

Da keine dieser Einflussfaktoren kontrolliert wird, lassen sich beobachtete Mittelwertveränderungen nicht eindeutig auf die Intervention zurückführen. Die Daten erlauben lediglich Aussagen darüber, dass sich Werte über die Zeit verändern – nicht jedoch, warum.

Die Problematik des Designs wird durch die Wahl der Datengrundlage weiter verschärft. Die Veränderungsmessungen basieren ausschließlich auf Selbsteinschätzungen der Schülerinnen und Schüler. Damit wird nicht die Entwicklung von Kompetenzen im Sinne tatsächlicher Leistungsfähigkeit erfasst, sondern die Veränderung subjektiver Einschätzungen.

Die Forschung zu Selbsturteilen zeigt, dass diese nur begrenzt mit objektiven Leistungsmaßen übereinstimmen und systematischen Verzerrungen unterliegen (Dunning et al. 2004; Paulhus / Vazire 2007). Besonders relevant ist in diesem Kontext,

dass die untersuchte Intervention selbst auf Reflexion, Selbstbewertung und die Verwendung von Kompetenzbegriffen abzielt.

Unter diesen Bedingungen ist zu erwarten, dass sich vor allem die Art und Weise verändert, wie Schülerinnen und Schüler über ihre Fähigkeiten urteilen. Steigende Werte in den Selbsteinschätzungen können daher ebenso gut Ausdruck erhöhter Sensibilität für Kompetenzbegriffe oder veränderter Bezugsnormen sein wie tatsächlicher Kompetenzzuwächse.

Die Messung erfasst somit unmittelbar eine Veränderung in den Selbsteinschätzungen, nicht notwendigerweise eine Veränderung der zugrunde liegenden Fähigkeiten.

Auch in Bezug auf diese Veränderungen treten nur geringe und inkonsistente Effekte auf. Die berichteten Effekte sind insgesamt klein, uneinheitlich oder statistisch unsicher. Effektstärken fallen gering aus, Konfidenzintervalle schließen teilweise Null ein, und Veränderungen zeigen kein konsistentes Muster über Skalen hinweg.

Solche Befunde erfordern eine entsprechend vorsichtige Interpretation. Kleine und instabile Effekte liefern keine belastbare Grundlage für weitreichende Schlussfolgerungen über die Wirksamkeit einer Intervention (Lakens 2013).

Im vorliegenden Beitrag wird diese Zurückhaltung nicht durchgehend eingehalten. Die Ergebnisse werden als Hinweise auf die „Förderung“ oder „Stärkung“ der 4K interpretiert, obwohl die Datenlage eher auf begrenzte und unsichere Veränderungen hindeutet. Ein zentrales Interpretationsproblem besteht somit in der systematischen Gleichsetzung von Veränderungen in Selbsteinschätzungen mit Veränderungen von Kompetenzen.

Diese Verschiebung ist methodisch nicht gerechtfertigt. Selbsteinschätzungen erlauben nur begrenzte Rückschlüsse auf die tatsächliche Leistungsfähigkeit. Ohne leistungsbasierte Messverfahren oder externe Kriterien bleibt unklar, ob beobachtete Veränderungen überhaupt eine inhaltliche Bedeutung haben.

Die Interpretation überschreitet damit die Aussagekraft der Daten. Aus Veränderungen in Selbsteinschätzungen werden Schlussfolgerungen über Kompetenzentwicklung gezogen, die empirisch nicht gedeckt sind.

Besondere Vorsicht ist schließlich beim Umgang mit schwachen oder nicht eindeutigen Befunden geboten. Wenn nicht signifikante oder kleine Effekte als „Tendenzen“ oder „erste Hinweise“ beschrieben werden, kann dies in einer explorativen Pilotstudie legitim sein. Zugleich sollte transparent bleiben, dass solche Befunde keine starke empirische Evidenz für Wirksamkeit darstellen.

Eine ausgewogene Interpretation müsste daher positive, ausbleibende und gegenläufige Veränderungen gleichermaßen berücksichtigen. Dazu gehört auch die Möglichkeit, dass einzelne Projektmaßnahmen keine messbaren Effekte auf die erfassten Selbsteinschätzungen hatten oder dass das eingesetzte Instrument für entsprechende Veränderungen nicht hinreichend passend war.

In der Gesamtschau erlauben die vorliegenden Veränderungsmessungen keine belastbaren Aussagen über die Wirkung der untersuchten Intervention. Das Design schließt zentrale Bedrohungen der internen Validität nicht aus, die Datengrundlage erfasst keine Kompetenzen im engeren Sinne, und die berichteten Effekte sind klein und inkonsistent.

Die daraus gezogenen Schlussfolgerungen gehen insgesamt teilweise deutlich über das hinaus, was die Daten unmittelbar stützen. Empirisch zeigen sich Veränderungen in Selbsteinschätzungen unter spezifischen Bedingungen. Ob und in welchem Ausmaß diese Veränderungen mit der Entwicklung der zugrunde gelegten Kompetenzen einhergehen, lässt sich auf Basis der vorliegenden Daten jedoch nicht eindeutig bestimmen.

4.3 ZWISCHENFAZIT

Zusammenfassend zeigt sich, dass das Forschungsdesign von Albrecht et al. (2025) strukturell stark auf ein bereits vorausgesetztes Kompetenzmodell ausgerichtet ist. Problemformulierung, Design, Datengrundlage und statistische Auswertung greifen dabei eng ineinander und orientieren sich durchgängig an diesem Modell. Diese enge Kopplung begrenzt die Möglichkeit, zentrale Annahmen unabhängig zu prüfen oder auch in Frage zu stellen. Empirische Verfahren tragen in diesem Rahmen primär dazu bei, das Modell differenziert auszuarbeiten und seine Anwendbarkeit im Untersuchungskontext zu beleuchten. Demgegenüber tritt ihre Funktion als Instrumente einer konsequent ergebnisoffenen Prüfung in den Hintergrund.

Damit liegt der zentrale Beitrag der Studie weniger in einem belastbaren Nachweis der Struktur und Förderbarkeit der 4K als in der Sichtbarmachung jener Herausforderungen, die mit der empirischen Operationalisierung eines breit angelegten Kompetenzrahmens verbunden sind. Die folgenden Analysen knüpfen daran an und prüfen genauer, welche Struktur sich in den vorliegenden Daten tatsächlich rekonstruieren lässt und wie die berichteten Befunde im Verhältnis zum formulierten Validierungsanspruch zu bewerten sind.

5 — RE-ANALYSE DER DATEN

Die im Folgenden berichteten Re-Analysen wurden durch die transparente Bereitstellung der Materialien von Albrecht et al. (2025) ermöglicht. Die Autorinnen und Autoren machen ergänzende methodische Informationen, Daten und Auswertungsskripte öffentlich zugänglich und schaffen damit eine Grundlage für Nachvollziehbarkeit, Replikation und weiterführende wissenschaftliche Auseinandersetzung (Bauer et al. 2026; Nosek et al. 2015), was gerade im Kontext eines methodisch anspruchsvollen und bildungspolitisch relevanten Themas besonders wertvoll ist. Sie erlaubt es der wissenschaftlichen Community, die Analysen nicht nur nachzuvollziehen, sondern auch alternative Modellierungen zu prüfen und zu explorieren. Die folgenden Analysen verstehen sich daher nicht als Infragestellung dieser Offenlegungspraxis, sondern als deren produktive Nutzung. Entsprechend stellen auch wir das vollstän-

dige Analyseskript unserer Re-Analysen in einem eigenen OSF-Repository bereit: <https://osf.io/35qnp/>

Damit soll sichergestellt werden, dass auch die hier berichteten Befunde nachvollziehbar, überprüfbar und weiterdiskutierbar sind.

Inhaltlich knüpfen unsere Analysen an den von Albrecht et al. formulierten Validierungsanspruch an und prüfen, inwiefern sich die postulierte Struktur der 4K in den bereitgestellten Daten rekonstruieren lässt. Im Zentrum stehen zwei Fragen: Erstens, ob die konfirmatorischen Modelle die angenommene Struktur der vier Kompetenzbereiche und ihrer Subdimensionen hinreichend stützen; zweitens, ob explorative Analysen alternative Strukturmuster nahelegen. Die Re-Analyse konzentriert sich damit auf die quantitativen Schülerdaten sowie auf die Frage der Konstruktstruktur; dies ist selbstverständlich zu unterscheiden von Aussagen zur praktischen Bedeutung des Modellprojekts, zu den qualitativen Befunden oder zur schulischen Innovationsleistung.

5.1 REPLIKATION DER KONFIRMATORISCHEN FAKTORENANALYSEN

Zunächst wurden die im Beitrag von Albrecht et al. (2025) berichteten konfirmatorischen Faktorenanalysen (CFA) auf Basis der bereitgestellten Daten und des Auswertungsskripts repliziert. Modelliert wurde für alle vier Kompetenzbereiche jeweils eine identische dreigliedrige Binnenstruktur mit den Faktoren Wissen/Fähigkeiten, Haltung und Gelegenheiten. Diese Struktur wird im Modell vorausgesetzt, ohne dass ihre empirische Tragfähigkeit zuvor unabhängig geprüft wird.

Die Replikation zeigt, dass dieses Messmodell keine empirische Bestätigung erfährt. Für das Modell mit einem Messzeitpunkt ergibt sich zwar ein RMSEA im akzeptablen Bereich (robust RMSEA = .050), die inkrementellen Fit-Indizes bleiben jedoch deutlich unter den in der Literatur üblichen Mindestwerten (robust CFI = .818; robust TLI = .802). Damit wird ein zentrales Kriterium konfirmatorischer Modellprüfung nicht erfüllt.

Im längsschnittlichen Modell über zwei Messzeitpunkte hinweg verschlechtert sich die Passung erheblich (robust CFI = .591; robust TLI = .575), was auf eine gravierende Missspezifikation des Modells hinweist. Hinzu kommt, dass die Re-Analyse Hinweise auf eine fehlende Trennbarkeit der Faktoren liefert, etwa in Form extrem hoher latenter Korrelationen bis in den Bereich von $r \geq 1$. Dies spricht gegen eine Interpretation der postulierten Dimensionen als klar voneinander abgrenzbare latente Konstrukte.

Die konfirmatorischen Analysen liefern damit keine hinreichende Evidenz für die angenommene Modellstruktur. Zentrale Modellannahmen, insbesondere die klare Trennbarkeit der postulierten Dimensionen, werden durch die vorliegenden Daten nicht gestützt. Dies gilt sowohl im Querschnitt als auch in der längsschnittlichen Modellierung. Die Befunde sprechen daher weniger für eine Bestätigung des Messmodells als für substanzielle Strukturprobleme und weiteren Klärungsbedarf.

5.2 EXPLORATIVE FAKTORENANALYSEN

Vor dem Hintergrund dieser Befunde prüften wir anschließend, welche latente Struktur sich in den Daten ohne Vorgabe des ursprünglichen Messmodells abzeichnet. Zur Klärung dieser Frage wurden explorative Faktorenanalysen durchgeführt, um die latente Struktur der Daten ohne Vorgabe eines theoretischen Modells zu untersuchen und zu prüfen, ob sich die postulierten Kompetenzbereiche und Subdimensionen auch unter explorativen Bedingungen rekonstruieren lassen.

Hierzu wurden bereichsspezifische EFAs für die vier Kompetenzbereiche sowie ergänzend bereichsübergreifende Analysen über alle Items gerechnet. Berücksichtigt wurden dabei insbesondere die inhaltliche Interpretierbarkeit der Faktoren, Hinweise auf Cross-Loadings und die Frage, ob die Daten entlang der theoretischen 4K-Aspekte oder anderer Itemcharakteristika wie etwa nach Selbsteinschätzungen, Haltungen oder wahrgenommenen Lerngelegenheiten strukturiert sind.

Grundlage der Analysen waren ordinale Items, die mittels polychorischer Korrelationen analysiert wurden. Die Faktorextraktion erfolgte mit der Minimum-Residual-Methode (minres) bei obliquer Rotation (oblimin). Die Bestimmung der Faktorzahl erfolgte anhand von Scree-Plot, Parallelanalyse und MAP-Test sowie unter Berücksichtigung der inhaltlichen Interpretierbarkeit. Für die Faktorzuordnung wurden Schwellenwerte definiert (Hauptladung $\geq .40$; problematische Cross-Loadings ab .30 bzw. bei geringer Differenz zwischen Haupt- und Nebenladung). Ergänzend wurden Sensitivitätsanalysen durchgeführt (klassenzentrierte Daten, metrische Modellierung, Leave-one-class-out).

Die bereichsspezifischen Analysen zeigen kein konsistentes Faktorenmuster über die vier Kompetenzbereiche hinweg. Für Kreativität und kritisches Denken ergeben sich mehrdimensionale Lösungen (zwei bis drei Faktoren), die sich als Trennung zwischen Kompetenzaspekten, wahrgenommenen Lerngelegenheiten und teilweise attitudinalen Komponenten interpretieren lassen. Für Kommunikation ergibt sich hingegen überwiegend ein Generalfaktor, während Kollaboration am plausibelsten durch eine zweidimensionale Struktur beschrieben wird.

Damit wird zugleich die zentrale Annahme eines einheitlichen Binnenmodells geprüft, das die Kompetenzbereiche in die Subdimensionen Wissen, Fähigkeiten/Fertigkeiten sowie Einstellungen/Werte/Haltungen untergliedert. Die explorativen Analysen liefern hierfür keine stabile empirische Grundlage. Zwar lassen sich in einzelnen Bereichen Faktoren identifizieren, die inhaltlich an diese Unterscheidung anschließen, eine konsistente und bereichsübergreifend replizierbare Differenzierung der drei Subdimensionen zeigt sich jedoch nicht. Insbesondere treten Wissens- und Fähigkeitsaspekte wiederholt gemeinsam auf und lassen sich empirisch nicht trennscharf unterscheiden. Attitudinale Komponenten zeigen ein uneinheitliches Muster, teilweise bilden sie eigene Faktoren, teilweise verschmelzen sie mit anderen Dimensionen oder treten nicht als eigenständige Dimension hervor. Die postulierte dreigliedrige Binnenstruktur lässt sich daher nicht als replizierbare latente Struktur der Daten rekonstruieren, sondern allenfalls in einzelnen Teilmustern anschlussfähig.

Über die einzelnen Bereiche hinweg zeigt sich jedoch ein robusteres Muster: Die Struktur folgt primär der Unterscheidung zwischen Kompetenz- bzw. Selbstbeschreibungsisems und Gelegenheitsisems. Attitudinale Komponenten treten uneinheitlich auf und bilden keine stabile, eigenständige Dimension. Die Daten weisen damit durchaus Struktur auf, diese verläuft jedoch quer zur theoretischen Gliederung der 4K.

Ergänzend wurde eine explorative Faktorenanalyse über alle Items hinweg durchgeführt, um zu prüfen, ob sich die vier Kompetenzbereiche selbst empirisch rekonstruieren lassen. Auch diese Analyse liefert keine Unterstützung für die postulierte 4K-Struktur. Die Bestimmung der Faktorzahl fällt uneinheitlich aus (Prätest: neun Faktoren nach Parallelanalyse/Scree-Plot, vier nach MAP; Posttest: acht bzw. sechs Faktoren), was bereits auf eine instabile Struktur hinweist.

Die Vier-Faktoren-Lösung im Prätest – als parsimonischste Lösung nach MAP – ordnet die Items entlang ihrer Funktionslogik: Ein dominanter Faktor bündelt Kompetenz- bzw. Selbstbeschreibungsisems („Ich kann/weiß“), ein zweiter Faktor erfasst wahrgenommene Lerngelegenheiten im Unterricht, während weitere Faktoren reflexive bzw. soziale Aspekte sowie kommunikative Austauschprozesse abbilden. Diese Struktur entspricht nicht der Differenzierung in die vier Kompetenzbereiche. Zusätzlich treten substantielle Cross-Loadings auf, sodass auch diese Lösung nur begrenzt trennscharf interpretierbar ist.

In der Zusammenschau der konfirmatorischen und explorativen Analysen ergibt sich ein konsistentes Bild: Weder auf Ebene der einzelnen Kompetenzbereiche noch in einer bereichsübergreifenden Betrachtung lässt sich die postulierte Struktur der 4K empirisch hinreichend rekonstruieren. Ebenso wenig lässt sich eine stabile Differenzierung der zugrunde gelegten Subdimensionen nachweisen. Die Annahme, dass es sich um eigenständige, trennbare Kompetenzdimensionen handelt, wird durch die vorliegenden Daten nicht gestützt. Vielmehr scheint ihre Organisation stärker durch Itemformate, Antwortlogiken und die Unterscheidung zwischen Selbstbeschreibungen und wahrgenommenen Gelegenheiten geprägt zu sein als durch die theoretische zugrunde gelegte 4(oder 3)K.

6 — EMPIRISCHE FORSCHUNG ALS BESTÄTIGUNGSINSTRUMENT: ZUR LOGIK DER STABILISIERUNG

Die faktorenanalytischen Ergebnisse stützen die postulierte Struktur der 4K nicht, und belastbare Hinweise auf ihre gezielte Förderung lassen sich aus den vorliegenden Daten nicht ableiten. Entscheidend ist jedoch nicht nur dieser Befund selbst, sondern die Art und Weise, wie mit ihm umgegangen wird.

Im Beitrag werden empirische Einschränkungen durchaus sichtbar gemacht: Die fehlende Trennbarkeit von Wissen und Fähigkeiten/Fertigkeiten wird berichtet, schwache und nicht signifikante Effekte werden benannt, und die Notwendigkeit weiterer Forschung wird anerkannt. Allerdings werden empirische Nicht-Bestätigungen nicht als solche interpretiert. Ergebnisse, die den ursprünglichen Modellannahmen nicht

entsprechen, führen nicht zu einer grundlegenden Relativierung der 4K, sondern werden anschlussfähig interpretiert.

Im Fall der Faktorenanalyse zeigt sich dies darin, dass die fehlende Trennbarkeit zentraler Dimensionen nicht als Widerlegung der Modellstruktur gewertet wird, sondern zur nachträglichen Zusammenlegung von Skalen führt. Ein Befund, der methodisch als Scheitern einer theoretischen Differenzierung zu interpretieren wäre, wird so in eine modifizierte Bestätigung des übergeordneten Konstrukts transformiert. Analog dazu werden in der Analyse der Veränderungsmessungen schwache oder unsichere Effekte als Hinweise auf Wirksamkeit gelesen, während alternative Erklärungen nicht systematisch kontrolliert werden können.

Dieses Muster verweist auf ein allgemeineres Spannungsverhältnis zwischen empirischer Forschung und normativ etablierten Kompetenzrahmen, das sich von einer klassischen hypothetisch-deduktiven Forschungslogik unterscheidet. In theorieprüfender Forschung fungieren empirische Daten als potenzielles Korrektiv: Sie müssen nicht nur erwartete Strukturen bestätigen können, sondern auch deren Einschränkung, Revision oder Zurückweisung ermöglichen. Diese Offenheit ist besonders wichtig, wenn die untersuchten Konzepte nicht primär aus wissenschaftlicher Theoriebildung hervorgegangen sind, sondern aus bildungspolitischen Zielsetzungen.

Der Beitrag von Albrecht et al. (2025) ist hierfür exemplarisch, weil er diese Spannung deutlich sichtbar macht. Einerseits leistet er einen transparent dokumentierten und für die weitere Forschung wertvollen Versuch, die 4K im schulischen Kontext empirisch zu erfassen. Andererseits zeigt gerade diese Offenlegung, dass die vorhandenen Daten die postulierte Struktur und die angenommenen Veränderungseffekte nur begrenzt tragen.

Eine solche Verschiebung ist insbesondere dann wahrscheinlich, wenn die untersuchten Konzepte nicht aus wissenschaftlicher Theoriebildung hervorgehen, sondern aus normativen und politischen Kontexten stammen. In diesen Fällen kann ein impliziter Erwartungsdruck entstehen, die Relevanz und Tragfähigkeit der Konzepte empirisch zu untermauern.

Empirische Forschung wird in solchen Konstellationen nicht nur zur Erkenntnisgewinnung eingesetzt, sondern steht zugleich in einem engen Verhältnis zu bereits etablierten Zielsetzungen. Für die Bewertung entsprechender Studien ergibt sich daraus ein veränderter Maßstab. Entscheidend ist nicht allein, ob einzelne methodische Entscheidungen vertretbar sind, sondern welche erkenntnistheoretische Funktion empirische Verfahren im Gesamtzusammenhang übernehmen. Dienen sie der offenen Prüfung eines Modells – einschließlich der Möglichkeit seiner Revision oder Zurückweisung? Oder tragen sie primär dazu bei, ein bereits normativ etabliertes Rahmenkonzept empirisch auszuarbeiten und zu stabilisieren?

Diese Frage stellt sich nicht nur für den hier diskutierten Beitrag, sondern verweist auf ein grundlegendes Spannungsverhältnis innerhalb der Bildungsforschung. Insbesondere bei Kompetenzmodellen, die aus bildungspolitischen Kontexten hervorgegangen sind, bewegt sich empirische Forschung häufig im Spannungsfeld zwischen wissenschaftlicher Prüfung und programmatischer Ausgestaltung.

Vor diesem Hintergrund ist zu klären, in welchem Maße empirische Verfahren zur kritischen Überprüfung solcher Modelle beitragen – und in welchem Maße sie zu deren wissenschaftlicher Ausformulierung und Festigung eingesetzt werden. Damit verschiebt sich der Status der präsentierten Ergebnisse von einer ergebnisoffenen empirischen Prüfung des 4K-Konzepts zu einer normativ vorbestimmten Ausarbeitung eines bereits vorausgesetzten Rahmens. Es stellt sich die Frage, in welchem Umfang empirische Verfahren hier tatsächlich als Instrumente kritischer Prüfung fungieren – oder in welchem Maße sie zugleich zur wissenschaftlichen Legitimierung bereits etablierter Annahmen beitragen.

7 — FAZIT

Die Analyse des Beitrags von Albrecht et al. (2025) zeigt ein konsistentes Spannungsverhältnis zwischen wissenschaftlichem Anspruch und empirischer Tragfähigkeit. Die Studie zielt darauf, Struktur und Förderbarkeit der 4K zu untersuchen, liefert dafür jedoch nur begrenzte Evidenz.

Empirisch erfasst werden Veränderungen in Selbsteinschätzungen im Rahmen eines Prä–Post-Designs ohne Kontrollgruppe. Diese lassen keine belastbaren Aussagen über die Entwicklung von Kompetenzen oder die Wirksamkeit der Intervention zu. Ebenso bleibt offen, ob die 4K und ihre Subdimensionen als trennbare Konstrukte empirisch angemessen abgebildet werden.

Damit ist weniger geklärt, ob das zugrunde liegende Modell empirisch trägt, als wie es unter den gegebenen Bedingungen operationalisiert werden kann. Die Ergebnisse werden überwiegend anschlussfähig an die Modellannahmen interpretiert, während deren mögliche Einschränkung oder Revision nur begrenzt in den Blick kommt.

Für die Bildungsforschung ergibt sich daraus die Notwendigkeit, normative Zielsetzungen und empirisch prüfbare Annahmen klarer zu unterscheiden und empirische Verfahren konsequent als Instrumente ergebnisoffener Prüfung zu nutzen.

LITERATURVERZEICHNIS

- **Albrecht, Christian et al. (2025)**: 21st Century Skills als Querschnittsaufgabe. Das Projekt „Zeitgemäße Lern- und Prüfungskultur entwickeln und erforschen“ (LUPE²) aus Perspektive des Deutschunterrichts. In: *MiDU – Medien im Deutschunterricht*, H. 2 (2025), 1–26. <https://doi.org/10.18716/ojs/midu/2025.2.1> — **American Educational Research Association / American Psychological Association / National Council on Measurement in Education (2014)**: *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association. — **Bauer, Johannes et al. (2026)**: *Empfehlungen zur Umsetzung von Open Science in der Empirischen Bildungsforschung – Gemeinsame Stellungnahme der GEBF und AEPF*. Frankfurt a.M.: Gesellschaft für Empirische Bildungsforschung. <https://doi.org/10.25656/01:34935> — **Bentler, Peter M. / Chou, Chih-Ping (1987)**: Practical Issues in Structural Modeling. In: *Sociological Methods & Research*, H. 1 (1987), 78–117. <https://doi.org/10.1177/0049124187016001004> — **Biesta, Gert J. J. (2007)**: Why „what works“ won't work: Evidence-based practice and the democratic deficit in educational research. In: *Educational Theory*, H. 1 (2007), 1–22. <https://doi.org/10.1111/j.1741-5446.2006.00241.x> — **Borsboom, Denny (2005)**: *Measuring the Mind. Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511490026> — **Brown, Timothy A. (2015)**: *Confirmatory Factor Analysis for Applied Research*. 2. Aufl. New York: Guilford Press. — **Campbell, Donald T. / Stanley, Julian C. (1963)**: *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin. — **DeVellis, Robert F. (2017)**: *Scale Development. Theory and Applications*. 4. Aufl. Thousand Oaks: Sage. — **Dunning, David / Heath, Chip / Suls, Jerry M. (2004)**: Flawed Self-Assessment. Implications for Health, Education, and the Workplace. In: *Psychological Science in the Public Interest*, H. 3 (2004), 69–106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x> — **Greß, Sotiria (2009)**: Governing by Numbers. The PISA „Effect“ in Europe. In: *Journal of Education Policy*, H. 1 (2009), 23–37. <https://doi.org/10.1080/02680930802412669> — **Kline, Rex B. (2016)**: *Principles and Practice of Structural Equation Modeling*. 4. Aufl. New York: Guilford Press. — **Lakens, Daniël (2013)**: Calculating and Reporting Effect Sizes to Facilitate Cumulative Science. In: *European Journal of Social Psychology*, H. 3 (2013), 261–269. <https://doi.org/10.1002/ejsp.1853> — **MacCallum, Robert C. / Roznowski, Mary / Necowitz, Lawrence B. (1992)**: Model Modifications in Covariance Structure Analysis. The Problem of Capitalization on Chance. In: *Psychological Bulletin*, H. 3 (1992), 490–504. <https://doi.org/10.1037/0033-2909.111.3.490> — **Nosek, Brian A. et al. (2015)**: Promoting an Open Research Culture. In: *Science*, 348 (6242), 1422–1425. <https://doi.org/10.1126/science.aab2374> — **OECD (2018)**: *The Future of Education and Skills. Education 2030*. Paris: OECD Publishing. <https://www.oecd.org/education/2030-project/> [19.06.2026]. — **Partnership for 21st Century Learning (2009)**: *Framework for 21st Century Learning*. <http://www.p21.org> [19.06.2026]. — **Patton, Michael Q. (2015)**: *Qualitative Research & Evaluation Methods*. 4. Aufl. Thousand Oaks: Sage. — **Paulhus, Delroy L. / Vazire, Simine (2007)**: The Self-Report Method. In: Robins, Richard W. / Fraley, R. Chris / Krueger, Robert F. (Hg.): *Handbook of Research Methods in Personality Psychology*. New York: Guilford Press, 224–239. — **Peng, Roger D. (2011)**: Reproducible Research in Computational Science. In: *Science*, 334 (6060), 1226–1227. <https://doi.org/10.1126/science.1213847> — **Popper, Karl R. (2002)**: *Logik der Forschung*. 11. Aufl. Tübingen: Mohr Siebeck. — **Rossi, Peter H. / Lipsey, Mark W. / Freeman, Howard E. (2004)**: *Evaluation. A Systematic Approach*. 7. Aufl. Thousand Oaks: Sage. — **Sellar, Sam / Lingard, Bob (2013)**: The OECD and the Expansion of PISA. New Global Modes of Governance in Education. In: *British Educational Research Journal*, H. 5 (2013), 710–725. <https://doi.org/10.1002/berj.3120> — **Shadish, William R. / Cook, Thomas D. / Campbell, Donald T. (2002)**: *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin. — **Weinert, Franz E. (2001)**: Concept of Competence. A Conceptual Clarification. In: Rychen, Dominique S. / Salganik, Laura H. (Hg.): *Defining and Selecting Key Competencies*. Seattle u.a.: Hogrefe & Huber, 45–65.

ÜBER DIE AUTOR*INNEN

Dr. habil. Benedikt Wisniewski ist zentraler Schulpsychologe an der Staatlichen Schulberatungsstelle für die Oberpfalz. Forschungsschwerpunkte umfassen Feedback im schulischen Kontext und Bildungsmythen.

Tanja Fritz ist wissenschaftliche Mitarbeiterin am Lehrstuhl Behavioral Health Technology der Universität Augsburg. Aktuelle Forschungsschwerpunkte umfassen innovative Methoden der Forschungssynthese sowie der Einsatz von Anreizen in Umfragestudien.

Prof. Dr. habil. Martin Daumiller leitet die Abteilung Pädagogische Psychologie an der Universität Freiburg. Aktuelle Forschungsschwerpunkte umfassen Motivation, Feedback, akademische Freiheit, KI in Bildungskontexten sowie Lehr-Lern-Prozesse in Schule und Hochschule.