
METHODISCHE KRITIK UND IHRE REICHWEITE: EINE ANTWORT AUF WISNIEWSKI, FRITZ & DAUMILLER (2026)

Christian Albrecht

Friedrich-Alexander-Universität Erlangen | christian.albrecht@fau.de

Jule Böhmer

Universität Hamburg | jule.boehmer@uni-hamburg.de

Stefanie Claußen

Pädagogische Hochschule Weingarten | stefanie.claussen@ph-weingarten.de

Holger Kaboth

Goethe Universität Frankfurt am Main | kaboth@sport.uni-frankfurt.de

Marcus Penthin

Friedrich-Alexander-Universität Erlangen | marcus.penthin@fau.de

Sophia Römer

Friedrich-Alexander-Universität Erlangen | sophia.roemer@fau.de

Fabian Ruth

Universität Tübingen | fabian.ruth@uni-tuebingen.de

Alexandra Säbsch

Friedrich-Alexander-Universität Erlangen

VORWORT

Wir danken Wisniewski, Fritz und Daumiller (2026) für die intensive Auseinandersetzung mit unserem Beitrag und die Re-Analyse der offen bereitgestellten Daten und Materialien. Ihre Replik zeigt exemplarisch zweierlei: Sie verdeutlicht, welche Chancen Open Science für den wissenschaftlichen Diskurs eröffnen kann, zugleich macht sie sichtbar, welche Verantwortung mit der Nachnutzung offener Daten und Materialien verbunden ist.

Unsere Antwort unterscheidet deshalb zwischen zwei Ebenen: Erstens greifen wir berechnete methodische Einwände gegen den quantitativen Teil unserer Begleitforschung auf. Die Re-Analysen zeigen, dass die angenommene Binnenstruktur des Erhebungsinstruments durch die vorliegenden Daten nur eingeschränkt gestützt wird und weiterer psychometrischer Prüfung bedarf. Entsprechend präzisieren wir den Status des Instruments und die Reichweite der daraus gewonnenen Befunde. Wir begrüßen ausdrücklich die Möglichkeiten, die sich daraus für die Überarbeitung und Weiterentwicklung des Erhebungsinstruments ergeben.

Zweitens prüfen wir, inwieweit sich Schlussfolgerungen aus diesen Befunden für den Ausgangsbeitrag insgesamt ableiten lassen. Wir widersprechen der Replik dort, wo sie die Kritik an einem begrenzten quantitativen Studienteil auf die übergeordnete Forschungslogik und die empirische Tragfähigkeit der gesamten Mixed-Methods- und Implementationsstudie überträgt.

1 — EINWÄNDE UND PRÄZISIERUNGEN

Wisniewski, Fritz und Daumiller (2026) beschränken ihre Re-Analysen auf die quantitativen Anteile unserer Begleitforschung, die bei zwei von acht Forschungsfragen eine Rolle spielten: die Erhebung von Selbsteinschätzungen der Schüler:innen in Bezug auf die 4K und ihre jeweiligen Subdimensionen sowie die wahrgenommenen Gelegenheiten, im Unterrichtsprojekt zu kommunizieren, zu kollaborieren, kreativ zu arbeiten und kritisch zu denken.

Kritisiert werden in der Replik insbesondere die theoretische Unschärfe der 4K, die eingeschränkte empirische Trennbarkeit der Konstrukte, die Nutzung von Selbsteinschätzungen und die fehlende Kontrollgruppe. Die Autor:innen argumentieren, dass auf dieser Grundlage keine starken Aussagen über die Struktur und Förderbarkeit der 4K getroffen werden können. Am Beispiel unseres Artikels wird problematisiert, dass empirische Befunde teilweise eher zur Stabilisierung eines bereits vorausgesetzten Kompetenzmodells genutzt würden, statt dieses ergebnisoffen zu prüfen.

Diese Kritik ist für den quantitativen Studienteil methodisch in Teilen nachvollziehbar. Im Folgenden greifen wir berechnete Einwände auf und präzisieren sie bzw. unsere Aussagen. Ihre Reichweite für die gesamte Studie wird anschließend in Abschnitt 2 geprüft.

1.1. REPRODUZIERBARKEIT DER DATENGRUNDLAGE

Zu Recht weisen Wisniewski, Fritz und Daumiller (2026) darauf hin, dass Reproduzierbarkeit „ein zentrales Gütekriterium quantitativer Forschung“ sei und „Datenbereinigungs- und Transformationsschritte [...] vollständig dokumentiert und idealerweise skriptbasiert nachvollziehbar sein“ (Wisniewski / Fritz / Daumiller 2026, 5) müssten. Kritisiert werden in diesem Zusammenhang manuelle und „nicht dokumentierte Eingriffe“ (ebd.), die dazu führten, dass unsere Analysen nicht überprüfbar seien und ihren wissenschaftlichen Status als replizierbare Befunde verlören. Ausgangspunkt dieses Hinweises der Replik sind insgesamt 11 doppelt auftretende IDs (acht in MZP 1, drei in MZP 2), die auf Eingabefehler seitens der Teilnehmer:innen und auf eindeutig bestimmbare Mehrfachteilnahmen mit frühen Abbrüchen im Fragebogen zurückzuführen sind. In längsschnittlichen Datensätzen ist es keine Seltenheit, dass beim Zusammenfügen der Messzeitpunkte mehrfache Zuordnungen auftreten können. Manuelle Eingriffe sind hier nötig, da sich nicht immer automatisiert Lösungen finden lassen. Diese wurden jedoch nicht undokumentiert verändert, sondern in einer im OSF-Repositorium hinterlegten und von Beginn an öffentlich zugänglichen Exceldatei erfasst und im Skript durch ausführbare Syntaxbefehle dokumentiert. Die Formulierung im Skript, nach der diese Eingriffe händisch in der Datendatei erfolgt seien, ist korrekturbedürftig. Tatsächlich fanden aber keinerlei Manipulationen in der Rohdatendatei statt (Dokumentation unserer Studie siehe <https://doi.org/10.17605/OSF.IO/SVJFH>).

Die dokumentationsbezogene Unschärfe rechtfertigt somit Kritik an der Nachvollziehbarkeit, aber nicht den Schluss, die Analysen seien insgesamt nicht überprüfbar und verlören deshalb ihren Status als replizierbare Befunde.

1.2. MISSINTERPRETATION ZENTRALER MODELLWARNUNGEN

Die Replik kritisiert zudem den „Umgang mit statistischen Warnmeldungen“ (ebd., 5). Bei den Berechnungen trat zweimal die Meldung „covariance matrix of latent variables is not positive definite; use lavInspect(fit, „cov.lv“) to investigate“ (lav_object_post_check) auf. Die reine Inspektion der Kovarianzmatrix auf positive Werte war in der Tat eine Fehlinterpretation dieser Warnmeldung. In beiden Fällen bezog sich diese Meldung jedoch nicht auf die im Artikel berichteten Analysen zum Fach Deutsch ($N = 90$), sondern auf die beiden CFAs des Gesamtmodells zu MZP 1 bzw. MZP 2 mit allen 12 latenten Variablen (Wissen/Können, Einstellungen/Haltungen sowie die Wahrnehmung von Gelegenheiten zur Ausübung der 4K) der Gesamtstichprobe über alle Fächer hinweg. Die Analysen mit der Gesamtstichprobe ($N = 224$) wurden der Transparenz wegen auch im Skript mit berichtet, aber nicht im Artikel bearbeitet, da der Fokus auf dem Fach Deutsch lag. In den beiden CFAs traten drei sehr hohe latente Korrelationen auf (zwischen der Subdimension „Einstellung“ zum kritischen Denken und zur Kollaboration zu MZP 1: $r = .91$, zu MZP 2: $r = .99$; zwischen der Subdimension „Wissen/Können“ beim kritischen Denken und bei Kommunikation zu MZP 2: $r = .92$). Die angesprochene Warnmeldung kann ursächlich in den angegebenen hohen Korrelationen liegen. Gleichzeitig können auch die geringe Stichprobengröße

(Kline 2023, 50) in Kombination mit der Komplexität der Gesamtmodelle zur Instabilität der Schätzungen beigetragen haben, auch wenn dies die Warnmeldung nicht relativiert.

Das R-Skript wurde aufgrund der Anmerkungen aus der Replik präzisiert und um detaillierte Angaben zu den bearbeiteten Mehrfachzuordnungen ergänzt, die fehlerhafte Interpretation der Warnmeldung der nicht positiv definiten Kovarianzmatrix entfernt sowie um zusätzliche Analysen der Kovarianzmatrix der beiden oben angesprochenen CFAs (vgl. ebd., 50f.) ergänzt. Diese Skriptversion ist nunmehr als „v2“ im [OSF-Projekt](#) verfügbar.

1.3. POST-HOC-ZUSAMMENLEGUNG VON FAKTOREN UND STATUS DES ERHEBUNGSINSTRUMENTS

Die Replik kritisiert, dass der Originalbeitrag von einem „grundlegend validierten“ (Albrecht et al. 2025, 23) Messinstrument spricht und dass die Skalen Wissen und Fähigkeiten/Fertigkeiten post-hoc zusammengelegt wurden (Wisniewski / Fritz / Daumiller 2026, 6). Die hohen Korrelationen zwischen diesen Subdimensionen sprechen tatsächlich gegen eine klare empirische Trennbarkeit, was allerdings bereits im Originalbeitrag berichtet wurde (vgl. Albrecht et al. 2025, 13). Die von Wisniewski, Fritz und Daumiller (2026, 9) durchgeführten explorativen Faktorenanalysen stützen diesen Befund.

In der LuPe²-Studie wurden für die Gesamtstichprobe alternative First-order-Faktorenmodelle bezüglich der Subdimensionen getestet und im Skript berichtet. Die LRT-Tests ergaben für jeden der 4K-Bereiche, dass das dreifaktorielle Modell unter Berücksichtigung der Zusammenlegung der oben angesprochenen Subdimensionen – wie im Artikel berichtet – am geeignetsten zu den Daten passte¹, auch wenn diese relative Überlegenheit weder eine insgesamt gute Modellpassung² noch eine stabile Replizierbarkeit der gefundenen Struktur belegt.

Die Zusammenlegung der Subdimensionen „Wissen“ sowie „Fähigkeiten/Fertigkeiten“ stellt daher eine transparente Reaktion auf die Datenlage dar. Im Sinne von MacCallum, Roznowski und Necowitz (1992) müssen solche nachträglichen Modellanpassungen sehr vorsichtig interpretiert werden, weil sie hoch anfällig für stichprobenspezifische Effekte sind. In einer frühen Phase der Instrumentenentwicklung können sie jedoch wichtige Hinweise darauf geben, welche theoretischen Unterscheidungen für Schüler:innen empirisch tatsächlich nachvollziehbar sind und welche weiter geschärft werden müssen. Denn inhaltlich ist es durchaus plausibel, dass Schüler:innen Wissen und Können in ihren Selbsteinschätzungen nicht klar voneinander trennen. Aussagen wie „Ich weiß, wie ich etwas machen muss“ und „Ich kann etwas tun“ können aus subjektiver Perspektive sehr eng beieinanderliegen. Die Zusammenlegung ist letztlich nicht als Bestätigung des ursprünglich angenommenen Modells zu verstehen, sondern als vorläufige, empirisch begründete Modifikation, die in weiteren Studien überprüft werden muss. Die explorativen Faktorenanalysen

¹ zweifaktoriell: 1. Wissen/Fähig-/Fertigkeiten, Einstellungen/Werte/Haltungen 2. Gelegenheit
dreifaktoriell: 1. Wissen, Fähig-/Fertigkeiten; 2. Einstellungen/Werte/Haltungen, 3. Gelegenheit
vierfaktoriell: 1. Wissen, 2. Fähig-/Fertigkeiten; 3. Einstellungen/Werte/Haltungen, 4. Gelegenheit

² In Bezug auf die Modellpassung sei angemerkt, dass der globale Fit auch dadurch negativ beeinträchtigt sein kann, wenn sich zahlreiche kleine Missspezifikationen eines sonst akzeptablen lokalen Fits aufsummieren (vgl. Kline 2023; Saris et al. 2009).

der Replik (vgl. Wisniewski / Fritz / Daumiller 2026, 9) stützen den Befund, dass die Items zu „Wissen“ bzw. „Fähigkeiten/Fertigkeiten“ mit gewissen Einschränkungen einen Faktor bilden und darüber hinaus die Items zu den wahrgenommenen Lerngelegenheiten ein robustes Muster zeigen. Attitudinale Komponenten („Haltungen“) treten allerdings uneinheitlich auf und es stellt sich insgesamt keine stabile Dimension ein. Die fehlende Trennbarkeit spricht daher nicht zwingend gegen die Relevanz der 4K als Kompetenzbereiche, sondern möglicherweise gegen eine zu feine Untergliederung einzelner Subdimensionen im eingesetzten Fragebogen.

Die Aussagekraft der vorliegenden Befunde zur Konstruktvalidität wird ebenfalls durch die Größe und heterogene Zusammensetzung der Stichprobe begrenzt. Diese setzt sich aus Schüler:innen unterschiedlicher Jahrgangsstufen und unterschiedlicher Fächer zusammen, die an unterschiedlichen Lehr-/Lernsettings teilgenommen haben. Mögliche Gruppenunterschiede könnten daher die ermittelte Faktorstruktur beeinflusst haben. Diese Einschränkung gilt gleichermaßen für die von Wisniewski, Fritz und Daumiller (2026) vorgenommenen Re-Analysen, da diese auf derselben Datengrundlage beruhen. Ihre explorativen Faktorenanalysen liefern daher wichtige zusätzliche Hinweise auf mögliche Strukturprobleme, können angesichts der begrenzten und heterogenen Stichprobe aber ebenso wenig als Grundlage für abschließende Aussagen zur Konstruktstruktur verstanden werden. Für zukünftige Untersuchungen wären eine größere Stichprobe und kontrolliertere Bedingungen wünschenswert.

Die eingeschränkte Trennschärfe einzelner Faktoren und Subdimensionen verweist darüber hinaus auf eine konzeptionelle Herausforderung der empirischen Forschung zu den 4K, die sich hier in einem theoretisch-methodischen Spannungsverhältnis befindet: Denn einerseits verlangt die quantitative Auseinandersetzung mit diesen 21st Century Skills eine trennscharfe Operationalisierung, will man Selbsteinschätzungen von Lernenden zu den einzelnen Dimensionen erheben, andererseits sind sie, wie bereits im Ausgangsbeitrag erwähnt, „interdependent zu verstehen“ (Albrecht et al. 2025, 6): Kollaboration basiert in der Regel auf Kommunikation, kreatives Arbeiten und Denken kann Anteile kritischen Denkens beinhalten und sich in ko-konstruktiven Prozessen entfalten, während kollaborative Lernprozesse nicht selten einen kreativen und kritischen Umgang mit dem Lerngegenstand erfordern. Auch Thornhill-Miller et al. (2023, 22) betonen die dynamischen Wechselbeziehungen zwischen den 4K und fassen sie konzeptionell als „Crea-Critical-Collab-ication“. Eine zusätzliche Herausforderung für die fächerübergreifende Operationalisierung besteht darin, dass die 4K zugleich domänenspezifische Facetten aufweisen. Die empirische Klärung des Verhältnisses zwischen diesen Anteilen zählt zu den von Binkley et al. (2012, 34) benannten „deep research challenges“.

Im Bewusstsein dieses Spannungsverhältnisses orientierten wir uns bei der Operationalisierung an den Ausarbeitungen von Binkley et al. (2012) und Thornhill-Miller et al. (2023). Diese konzeptionelle Grundlage bleibt in der Replik unberücksichtigt, obwohl das Framework von Binkley et al. ausdrücklich als Modell für die Entwicklung von Erhebungen von 21st Century Skills angelegt ist (vgl. Binkley et al. 2012, 33). Dass sie sich unter anderem auch auf bildungspolitische Diskurse beziehen, sagt zunächst nichts über ihre wissenschaftliche Untersuchbarkeit aus. Bildungspolitisch

geprägte Kompetenzmodelle sind in der Bildungsforschung nicht ungewöhnlich. Entscheidend ist vielmehr, ob sie konzeptionell präzisiert, transparent operationalisiert und empirisch ergebnisoffen geprüft werden. Das Meta-Modell von Binkley et al. basiert auf internationalen Lehrplan- und Frameworkanalysen und nimmt vornehmlich eine erziehungswissenschaftliche Perspektive mit Fokus auf schulische Bildung ein (vgl. Ehlers et al. 2024). So erschien uns dieses Framework für unsere Instrumentenentwicklung trotz seiner begrenzten empirischen Absicherung in Anschluss an Seidl (2024, 493) als „the best-backed model available“.

Vor diesem Hintergrund ist unser Beitrag explorativ zu verstehen, indem er zeigt, wie ein bildungspraktisch relevantes, konzeptionell und empirisch jedoch nicht abschließend geklärtes Modell in schulischen Kontexten operationalisiert und implementiert wird und welche Grenzen dabei sichtbar werden. Die Bezeichnung des Fragebogens als „grundlegend validiertes Messinstrument“ halten wir rückblickend für zu weitreichend. Gerne folgen wir dem Vorschlag der Replik, dass unsere Ergebnisse „als ein früher Schritt der Instrumentenentwicklung zu verstehen [sind], der Hinweise auf mögliche Skalenbildung liefert, zugleich aber substantielle Fragen der Konstruktvalidität offenlässt“ (ebd., 6).

Anschlussstudien sollten folglich nicht nur die angenommene Faktorenstruktur der 4K untersuchen, sondern auch alternative Itemformulierungen und Modellierungen prüfen. Denkbar wären insbesondere Modelle, welche die theoretisch angenommene Interdependenz der 4K abbilden, sowie hierarchische Modelle, die zwischen übergreifenden und domänenspezifischen Anteilen unterscheiden.

2 — ERKENNTNISANSPRUCH DER STUDIE UND GELTUNGSBEREICH DER KRITIK

Die Replik bietet zugleich Anlass, die Verantwortung zu reflektieren, die mit der wissenschaftlichen Nachnutzung offen bereitgestellter Daten verbunden ist. Offene Daten ermöglichen Kritik und Nachprüfung, verlangen aber zugleich eine sorgfältige und transparente Einordnung der Reichweite der eigenen Schlüsse. Im Umgang mit ihnen ist es daher wichtig, den geprüften Ausschnitt und den Geltungsanspruch der daraus abgeleiteten Kritik deutlich auszuweisen. Ebenso gehört dazu, den Erkenntnisanspruch des Ausgangsbeitrags anhand seiner Forschungsfragen, seines Designs und seiner Ergebnisformulierungen präzise zu rekonstruieren und ihn nicht an Ansprüchen zu messen, die er selbst nicht erhebt und die sich aus diesem Grund auch nicht von ihm ableiten lassen. Dieser Verantwortung werden Wisniewski, Fritz und Daumiller (2026) trotz ihrer konstruktiven Hinweise nicht durchgehend gerecht.

2.1. PILOTCHARAKTER DER LUPE²-STUDIE

Dies betrifft erstens die Formulierung und Einordnung des Anspruchs der kritisierten Studie. Die Replik legt die Lesart nahe, als wolle der zugrunde liegende Beitrag vorrangig die Struktur und Förderbarkeit der 4K empirisch belegen. Zu wenig berücksichtigt sie dabei, dass es sich bei der LuPe²-Studie um eine Pilotstudie im Rahmen

der Begleit- und Implementationsforschung handelt, was bereits im Originalbeitrag deutlich wird (vgl. Albrecht et al. 2025, 7, 19, 23).

Methodisch ist dies relevant, weil Pilotstudien gerade nicht primär dazu dienen, endgültige Wirksamkeitsnachweise zu liefern. Ihr Wert liegt vielmehr darin, Verfahren zu erproben, Machbarkeit zu prüfen, Instrumente und Forschungsdesigns weiterzuentwickeln und auf der Grundlage erster Erfahrungen im Feld konzeptionelle und methodische Präzisierungen für weiterführende Untersuchungen vorzunehmen (vgl. König 2018; Leon / Davis / Kraemer 2011; van Teijlingen/Hundley 2001).

Unter dieser Perspektive verschiebt sich der Bewertungsmaßstab: Auch wenn der Pilotcharakter nicht von der psychometrischen Prüfung der eingesetzten Instrumente entbindet, geht es nicht darum, ein neues Kompetenzmodell zu entwickeln oder kausale Effekte nachzuweisen, sondern darum, erste deskriptive und explorative Hinweise zur schulischen Umsetzung, zu Wahrnehmungen der Schüler:innen, zu Lehrkräfteperspektiven und zu Herausforderungen einer veränderten Lern- und Prüfungskultur zu gewinnen.

2.2. PSYCHOMETRISCHE KRITIK IN EINER MIXED-METHODS-STUDIE

Zweitens berücksichtigt die Replik den Mixed-Methods-Charakter des Originalbeitrags nicht hinreichend, soweit sie ihre Kritik am quantitativen Studienteil auf die übergeordnete Forschungslogik des Gesamtbeitrags überträgt. In der LuPe²-Studie werden quantitative Schüler:innendaten mit qualitativen Lehrkräftedaten ins Verhältnis gesetzt. In Bezug auf die Daten der Schüler:innen formulieren wir: „Darüber hinaus bieten die quantitativen Daten einen ersten Einblick in die Einstellungen und Haltungen der Schüler:innenschaft vor und nach der Projektdurchführung und geben Aufschluss darüber, inwiefern die in den Expert:inneninterviews erläuterten innovativen Lehr-/Lernformate überhaupt wahrgenommen werden.“ (Albrecht et al. 2025, 20). In der Mixed-Methods-Forschung wird gerade diese Verbindung als Stärke verstanden, weil komplexe pädagogische Phänomene selten durch eine einzige Datenquelle angemessen erfasst werden können (vgl. Johnson / Onwuegbuzie 2004). Die Integration qualitativer und quantitativer Daten relativiert nicht die psychometrischen Grenzen des quantitativen Studienteils, erweitert aber die Perspektive um die schulische Implementation. Die Replik weist zwar an einer Stelle darauf hin, dass sich ihre Re-Analyse auf die quantitativen Schüler:innendaten konzentriert und von Aussagen zur praktischen Bedeutung des Modellprojekts sowie zu den qualitativen Befunden zu unterscheiden ist (vgl. Wisniewski / Fritz / Daumiller 2026, 8). Gleichwohl werden auf diese Weise zentrale Fragestellungen und das damit verbundene qualitative Forschungsdesign ausgespart: berufsbezogene Überzeugungen (Forschungsfrage 1), die Integration der 4K in Unterrichts- und Prüfungskultur und in diesem Zusammenhang wahrgenommene Chancen und Herausforderungen (Forschungsfrage 2.1, 2.2 sowie 3.1 und 3.2) sowie die Identifikation von Ausprägungen und Wirkweisen der Kultur der Digitalität in der Projektarbeit der Lehrkräfte (Forschungsfrage 4). Natürlich ist es im Rahmen einer Replik legitim, nur Ausschnitte einer Studie zu prüfen – im vorliegenden Fall also zwei der acht Forschungsfragen. Wenn sie daraus jedoch

Schlussfolgerungen über die „übergeordnete Forschungslogik“ ableitet (Wisniewski / Fritz / Daumiller 2026, 2), müsste sie darlegen, inwiefern diese auch die qualitativ bearbeiteten Forschungsfragen und die Integration beider Studienteile betreffen.

2.3. ZUGESCHRIEBENER WIRKSAMKEITSANSPRUCH UND TATSÄCHLICHES ERKENNTNISINTERESSE

Drittens rekonstruiert die Replik den Erkenntnisanspruch unserer Studie an mehreren Stellen stärker, als er im Beitrag selbst formuliert ist. Sie legt wiederholt eine Lesart nahe, nach der unser Beitrag primär darauf angelegt sei, die Wirksamkeit von LuPe² oder die Förderbarkeit der 4K kausal anhand der quantitativen Daten nachzuweisen. Dies ist nicht der Fall. Ein kausaler Wirksamkeitsnachweis war weder Ziel noch Gegenstand der Studie. Die im Zusammenhang mit den Schüler:innendaten einmalige Verwendung des Begriffs „Auswirkungen“ (Albrecht et al. 2025, 19) in der Einleitung des Diskussionsteils mag missverständlich wirken, begründet jedoch keine kausalen Zuschreibungen. Forschungsfrage 5b richtet sich ausschließlich auf Unterschiede in den Selbsteinschätzungen vor und nach den Projekten. Diese werden an keiner Stelle als kausal nachgewiesene Effekte ausgewiesen. Unser Erkenntnisinteresse richtet sich auf die Implementation einer 4K-orientierten Lern- und Prüfungskultur unter realen schulischen Bedingungen, nicht auf die bloße Affirmation eines vorab gesetzten Modells. Wir haben untersucht, wie ein Netzwerk aus Praktiker:innen dieses Modell als Orientierungsrahmen für Unterrichts- und Prüfungskulturentwicklung implementiert, welche Herausforderungen und Gelingensbedingungen bei der schulischen Implementation der 4K antizipiert werden und welche Hinweise die Selbsteinschätzungen der Schüler:innen darauf geben, inwiefern sie die im Projekt geschaffenen Lerngelegenheiten wahrnahmen. Dies spiegelt sich in den aufgeführten acht Forschungsfragen wider.

Wir teilen die Einschätzung, dass Begleit- und Implementationsforschung aufgrund ihrer Nähe zum Feld, der ko-konstruktiven Entwicklung pädagogischer Innovationen und der normativen Zielorientierung besonders reflexiv mit dem Risiko affirmativer Ergebnisinterpretationen umgehen muss. Wir stimmen ebenso darin mit den Autor:innen der Replik überein, dass die Überführung normativer Zielkategorien in wissenschaftliche Forschung nicht trivial ist. Diesen Risiken ist durch eine transparente Ausweisung von Limitationen, methodische Triangulation, eine ausdrückliche Berücksichtigung nicht intendierter Befunde sowie durch eine Begrenzung der Ergebnisansprüche zu begegnen. Gleichzeitig zielt Evaluationsforschung nach Rossi, Lipsey und Freeman (2004) nicht ausschließlich auf den Nachweis isolierter Effekte. Sie untersucht auch Forschungslogik, Umsetzungsbedingungen, Prozessqualität und Kontextfaktoren. Genau darin liegt ein wesentlicher Erkenntniswert des Originalbeitrags: Er macht sichtbar, wie Lehrkräfte 4K-orientierte Lern- und Prüfungsformate im schulischen Alltag aufgreifen, wo Umsetzungsmöglichkeiten bestehen und an welchen Stellen traditionelle Prüfungskulturen an Grenzen geraten.

Der Vorwurf, uns ginge es darum, „die Plausibilität eines vorab gesetzten Reformanliegens zu stützen“ (Wisniewski / Fritz / Daumiller 2026, 4), ist letztlich auch des-

halb nicht überzeugend, da der Beitrag die Erwartungen an eine erfolgreiche Implementation keineswegs ungebrochen bestätigt: Die meisten Veränderungen in den Selbsteinschätzungen fallen schwach oder vernachlässigbar aus, die Einstellungen zur Kommunikation sind rückläufig und bei mehreren Skalen werden Messprobleme sowie Grenzen der Interpretation benannt. Darüber hinaus weist der Originalbeitrag ausdrücklich auf die kriteriengeleitete Auswahl der beteiligten Schulen und Lehrkräfte, die unterschiedlichen Förderansätze, die uneinheitliche Dauer der Projekte und die geringe Kontrolle über die Erhebungsbedingungen hin (vgl. Albrecht et al. 2025, 22f.).

2.4. ZUGESCHRIEBENE GLEICHSETZUNGEN

Die Kritik von Wisniewski, Fritz und Daumiller (2026) an einer „[s]ystematischen Gleichsetzung von Veränderungen in Selbsteinschätzungen mit Veränderungen von Kompetenzen“ (Wisniewski, Fritz & Daumiller 2026, 7) bezieht sich viertens auf eine Argumentation, die im Beitrag so nicht vertreten wird. Zutreffend ist, dass Selbsteinschätzungen keine objektiven Kompetenzmessungen darstellen und nur begrenzte Rückschlüsse auf tatsächliche Leistungsfähigkeit erlauben. Genau deshalb bezeichnet der Originalbeitrag bei der Darstellung der quantitativen Schüler:innendaten die erfassten Größen ausdrücklich als „Selbsteinschätzungen“, „selbst eingeschätzte Kompetenzen“ und „selbst eingeschätzte Wissens- und Kompetenzzuwächse“ (Albrecht et al. 2025, 21f.). Zudem wird ausdrücklich zwischen diesen subjektiven Angaben und künftig ergänzend zu erhebenden objektiven Kompetenzdaten unterschieden (vgl. ebd., 23). Die Studie ist damit nicht als Messung objektiver Kompetenzentwicklung angelegt, sondern als Erhebung subjektiver Wahrnehmungen, Einstellungen und wahrgenommener Lerngelegenheiten. Hierfür sind Selbstauskünfte eine grundsätzlich angemessene Datenquelle (vgl. Paulhus & Vazire 2007).

Auch die in der Replik ohne nähere Seitenangabe zitierten Begriffe „Stärkung“ und „Förderung“ (Wisniewski / Fritz / Daumiller 2026, 7) der 4K belegen die behauptete Gleichsetzung nicht. Wisniewski, Fritz und Daumiller (2026) verwenden sie, ohne ihren Kontext zu berücksichtigen, da sich die Begriffe nicht auf die quantitativen Selbsteinschätzungen der Schüler:innen im Fragebogen beziehen, sondern in einer Gesamtschau auf die qualitative Rekonstruktion der Einschätzungen der befragten Lehrpersonen und der berichteten Unterrichtsmaßnahmen (vgl. Albrecht et al. 2025, 20ff.).

2.5. AUSGEWIESENE ERKENNTNISGRENZEN

Die Replik weist fünftens darauf hin, dass ohne Kontroll- oder Vergleichsgruppe keine belastbaren kausalen Aussagen möglich sind. Der Hinweis ist richtig, jedoch haben wir einen solchen Anspruch weder erhoben noch führen wir die beobachteten Veränderungen kausal auf die Intervention zurück. Das Design erlaubt keine kausale Attribution der beobachteten Unterschiede, ermöglicht aber deskriptive Prä-Post-Vergleiche. Auf dieser Grundlage lässt sich beschreiben, welche Unterschiede in den Selbsteinschätzungen im Projektverlauf sichtbar wurden und welche Hinweise sich daraus für Anschlussstudien ergeben.

In diesem Zusammenhang zeigt sich ein Problem der Darstellung: Die Replik erläutert methodische Einschränkungen, die für die Einordnung der Befunde zweifellos relevant sind, macht aber nicht hinreichend deutlich, dass diese Einschränkungen von uns selbst benannt werden. Durch den Verzicht auf diesen Hinweis erhalten die grundlegenden Erläuterungen von Wisniewski und Kolleg:innen hier und immer wieder auch an anderer Stelle eine rhetorische Funktion: Sie vermögen den Eindruck zu erzeugen, dass wir uns dieser methodischen Grundlagen nicht bewusst wären. Der Originalbeitrag formuliert die Einschränkungen durch die fehlende Kontrollgruppe zwar nicht in derselben Ausführlichkeit, weist jedoch ausdrücklich auf die geringe Kontrolle über die Interventionsbedingungen hin und bezeichnet den Vergleich mit Klassen des Regelunterrichts als wichtigen Schritt für künftige Aussagen über die Effektivität der Interventionen (vgl. Albrecht et al. 2025, 22f.). So werden im Ausgangsbeitrag ausgewiesene Limitationen nachträglich von der Replik in Versäumnisse umgedeutet.

3 — FAZIT

Zusammenfassend begrüßen wir die konstruktiven Hinweise und detaillierten Re-Analysen zur Faktorstruktur des Erhebungsinstruments. Sie helfen, die Reichweite unserer quantitativen Befunde genauer einzuordnen, das Instrument angesichts der Herausforderungen bei der Operationalisierung der Konstrukte weiterzuentwickeln und künftige Studien methodisch weiter zu fundieren.

Problematisch ist die Replik jedoch dort, wo sie die Kritik an einzelnen quantitativen Analysen zu der Diagnose verallgemeinert, die Studie zeige insgesamt ein „konsistentes Spannungsverhältnis zwischen wissenschaftlichem Anspruch und empirischer Tragfähigkeit“ (Wisniewski / Fritz / Daumiller 2026, 11). Die Entwicklung und erste psychometrische Prüfung des Erhebungsinstruments bildeten einen Teil der LuPe²-Studie; die Studie war jedoch weder als abschließende Prüfung der operationalisierten 4K-Modellstruktur noch als kontrollierter Nachweis kausaler Interventionseffekte angelegt. Ein wesentlicher Erkenntniswert liegt nicht zuletzt darin, die schulische Implementation einer 4K-orientierten Lern- und Prüfungskultur aus den Perspektiven der Lehrkräfte und Schüler:innen sowie im Hinblick auf Unterrichtsentwicklung und Prüfungskultur zu untersuchen. Diesen mehrperspektivischen Anspruch reduziert die Replik auf den quantitativ auswertbaren Ausschnitt.

Gerade im Kontext von Open Science ist eine solche Reduktion nicht trivial. Eine Replik darf selektiv prüfen, muss ihren Geltungsanspruch dann aber entsprechend begrenzen. Werden bereits ausgewiesene Limitationen als von außen aufzudeckende Versäumnisse dargestellt oder werden dem Ausgangsbeitrag Ansprüche zugeschrieben, die er in dieser Form nicht erhebt, verschiebt sich die produktive methodische Klärung zu einer pauschalen Bewertung der gesamten Studie, die durch den untersuchten Ausschnitt nicht hinreichend gedeckt ist.

Unsere Antwort versteht sich deshalb nicht als Zurückweisung methodischer Kritik, sondern als Präzisierung ihres Geltungsbereichs. Die Re-Analyse (vgl. ebd., 8ff.) zeigt, dass die Operationalisierung der 4K und die psychometrische Qualität des

Erhebungsinstruments weiterentwickelt werden müssen. Die im Titel der Replik gewählte Formulierung „Die 4K im Gewand empirischer Forschung“ ist in diesem Zusammenhang jedoch unangemessen polemisch: Eine solche Rahmung beschränkt sich nicht auf die sachliche methodische Kritik einzelner Analysen, sondern stellt undifferenziert den wissenschaftliche Status des Gesamtbeitrags infrage.

Die 4K als didaktische Querschnittsaufgabe schulischer Lern- und Prüfungskulturen bleiben letztlich ein legitimer und weiterhin offener Gegenstand empirischer Bildungsforschung. LuPe² hat dazu einen ersten Beitrag geleistet – ausdrücklich vorläufig, aber als Ausgangspunkt für die weitere Auseinandersetzung gedacht.

QUELLENVERZEICHNIS

- **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. DOI: 10.18716/ojs/midu/2025.2.1. ——— **Binkley, Marilyn et al. (2012)**: Defining Twenty-First Century Skills. In: Griffin, Patrick / McGaw, Barry / Care, Esther (Hg.): *Assessment and Teaching of 21st Century Skills*. Dordrecht: Springer, 17-66. DOI: 10.1007/978-94-007-2324-5_2. ——— **Ehlers, Ulf-Daniel et al. (2024)**: Future Skills für Hochschulen. Eine kritische Bestandsaufnahme. In: Stifterverband für die Deutsche Wissenschaft (Hg.): *Future Skills lehren und lernen. Schlaglichter aus Hochschule, Schule und Weiterbildung*. Online-Vorveröffentlichung. <https://www.future-skills.net/analysen/vorveroeffentlichung-future-skills-fuer-hochschulen-eine-kritische-bestandsaufnahme> [06.08.2026]. ——— **Johnson, R. Burke / Onwuegbuzie, Anthony J. (2004)**: Mixed Methods Research. A Research Paradigm Whose Time Has Come. In: *Educational Researcher*, H. 7 (2004), 14-26. DOI: 10.3102/0013189X033007014. ——— **Kline, Rex B. (2023)**: *Principles and Practice of Structural Equation Modeling*. 5. Aufl. New York: Guilford. ——— **König, Lisa (2018)**: Planung und Vorbereitung empirischer Erhebungen. Oder: Der lange Weg zur eigentlichen Arbeit. In: Boelmann, Jan M. (Hg.): *Empirische Forschung in der Deutschdidaktik. Band 1: Grundlagen*. Baltmannsweiler: Schneider Hohengehren, 237-256. ——— **Leon, Andrew C. / Davis, Lori L. / Kraemer, Helena C. (2011)**: The Role and Interpretation of Pilot Studies in Clinical Research. In: *Journal of Psychiatric Research*, H. 5 (2011), 626-629. DOI: 10.1016/j.jpsychires.2010.10.008. ——— **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. DOI: 10.1037/0033-2909.111.3.490. ——— **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, 224-239. ——— **Rossi, Peter H. / Lipsey, Mark W. / Freeman, Howard E. (2004)**: Assessing and Monitoring Program Process. In: Rossi, Peter H. / Lipsey, Mark W. / Freeman, Howard E.: *Evaluation. A Systematic Approach*. 7. Aufl. Thousand Oaks: Sage, 169-201. ——— **Saris, Willem E. / Satorra, Albert / van der Veld, William M. (2009)**: Testing Structural Equation Models or Detection of Misspecifications? In: *Structural Equation Modeling*, H. 4 (2009), 561-582. DOI: 10.1080/10705510903203433. ——— **Seidl, Tobias (2024)**: Formative Assessment of 21st Century Skills. In: Ehlers, Ulf-Daniel / Eigbrecht, Laura (Hg.): *Creating the University of the Future. A Global View on Future Skills and Future Higher Education*. Wiesbaden: Springer VS, 489-507. DOI: 10.1007/978-3-658-42948-5_24. ——— **Thornhill-Miller, Branden et al. (2023)**: Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. In: *Journal of Intelligence*, H. 11, 1-32. ——— **van Teijlingen, Edwin R. / Hundley, Vanora (2001)**: The Importance of Pilot Studies. In: *Social Research Update*, H. 35, 1-4. ——— **Wisniewski, Benedikt / Fritz, Tanja M. / Daumiller, Martin (2026)**: Die 4K im Gewand empirischer Forschung. Eine Antwort auf Albrecht et al. (2025). In: *MiDU – Medien im Deutschunterricht*, H. 2 (2025), 1-20. DOI: 10.18716/OJS/MIDU/2025.2.6.

ÜBER DIE AUTOR*INNEN:

Dr. Christian Albrecht ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Didaktik der deutschen Sprache und Literatur an der FAU Erlangen-Nürnberg und unterrichtet Deutsch und Geographie an einem Nürnberger Gymnasium. Seine Forschungsschwerpunkte liegen in Fragen einer zeitgemäßen Prüfungskultur, digitaler Bildung, literarästhetischer Kommunikation, der Filmdidaktik sowie in der empirischen Unterrichtsforschung.

Dr. Jule Böhmer ist wissenschaftliche Mitarbeiterin an der Universität Hamburg für die Fachdidaktik der Einwanderungssprachen, Fachseminarleiterin für Russisch am Landesinstitut für Lehrerbildung und Schulentwicklung und unterrichtet Russisch und Geographie an einem Hamburger Gymnasium. Zu ihren Interessensschwerpunkten zählt die Gestaltung von Lern- und Prüfungsszenarien in der Kultur der Digitalität, insbesondere im Sprachunterricht.

Stefanie Claußen ist Referentin für Qualitätsmanagement an der Pädagogischen Hochschule Weingarten und war dort als akademische Mitarbeiterin im Projekt TE-goDi beteiligt an der Entwicklung digitaler Lehr-Lern-Angebote für das Lehramtsstudium. An der Hamburger Fern-Hochschule unterstützt sie die Entwicklungsabteilung Lehren und Lernen.

StR. Holger Kaboth ist wissenschaftlicher Mitarbeiter am Institut für Sportwissenschaften der Goethe-Universität Frankfurt. Er arbeitet zu schulischer Heterogenität, hybriden Lehr-Lernformaten und Schulentwicklung. Ein besonderer Fokus liegt auf der Förderung wissenschaftspropädeutischen Lernens sowie der Verbindung von Forschung und schulischer Praxis.

Dr. Marcus Penthin ist Dozent an der FAU Erlangen-Nürnberg. Als vorrangig quantitativ arbeitender empirischer Bildungsforscher beschäftigt er sich u.a. mit religiösen und berufsbezogenen Überzeugungen von Lehrkräften, Religiosität bei geflüchteten Jugendlichen und Digitalisierung in der kulturellen Bildung. Er ist in der Forschungs- und Methodenberatung tätig und ein Verfechter von Open Science.

Sophia Römer ist wissenschaftliche Mitarbeiterin am Institut für Grundschulforschung an der FAU Erlangen-Nürnberg. Zu ihren Schwerpunkten zählen neben zeitgemäßer Prüfungskultur Inklusion, professionelle Kompetenzen von Lehrkräften und Lehramtsstudierenden sowie Demokratiebildung im Grundschulunterricht.

Dr. Fabian Ruth ist wissenschaftlicher Mitarbeiter an der Forschungsstelle Präsentationskompetenz, angesiedelt am Forschungszentrum für Wissenschaftskommunikation der Universität Tübingen. Zu seinen Tätigkeiten und Forschungsschwerpunkten gehören die wissenschaftliche Evaluation der Bildungsinitiative „Jugend präsentiert“, die valide Bewertung von Präsentationen sowie die digitalgestützte Förderung kommunikativer Kompetenzen.

Alexandra Säbsch ist Studentin des Grundschullehramts sowie studentische Hilfskraft am Lehrstuhl für Didaktik der deutschen Sprache und Literatur an der FAU Erlangen-Nürnberg.