Guttke, Joel; Porsch, Raphaela

Inhaltsvalidierung von Fragebogenitems zu kognitiver Aktivierung im Englischunterricht der Grundschule mittels Expert:innenbefragung

Lohe, Viviane [Hrsg.]; Lindl, Alfred [Hrsg.]; Kirchhoff, Petra [Hrsg.]: Unterrichtsqualität in schulischen Fremdsprachen. Theoretische Ansätze und empirische Ergebnisse aus den Fachdidaktiken. Münster ; New York : Waxmann 2024, S. 57-84

Joel Guttke und Raphaela Porsch

Inhaltsvalidierung von Fragebogenitems zu kognitiver Aktivierung im Englischunterricht der Grundschule mittels Expert:innenbefragung

Abstract der Herausgeber:innen
In their article "Content validation of questionnaire items on cognitive activation in primary school English lessons using an expert survey," Raphaela Porsch and Joel Guttke focus on the role of the basic dimension of cognitive activation in English lessons. Up to now, the findings on this central dimension of instructional quality have been quite mixed. Therefore, this contribution takes a closer empirically founded look at this dimension. It presents the results of an online survey with selected experts on the content validity of theory-based constructed items to collect data on cognitive activation in fourth-grade English classes. In a next step, questionnaires for pupils and teachers were used to target the same construct from a different perspective. Implications for the operationalisation of the construct and the revision of items will then be derived. The approach described also highlights the added value of expert surveys in the process of constructing new instruments in research-based foreign language teaching.

1 Einleitung

Im deutschsprachigen Raum gelten die drei Basisdimensionen von Unterrichtsqualität als etabliertes Modell zur empirischen Erfassung von Unterrichtsmerkmalen und Erklärung der Lernwirksamkeit von Unterricht (Klieme, 2019). Im Gegensatz zu anderen Modellen der Unterrichtsqualität fußt das Modell hinsichtlich seiner faktoriellen Struktur auf einer vergleichsweise stabilen empirischen Basis. In Bezug auf die prädiktive Validität für Schüler:innenleistungen fällt die Befundlage jedoch – insbesondere für die Basisdimension der kognitiven Aktivierung – inkonsistent aus (z. B. Praetorius et al., 2018). Dennoch wird kognitiver Aktivierung auch in anderen Modellen der Unterrichtsqualität weiterhin eine zentrale Rolle zugeschrieben (Lindl et al., in diesem Band). Die Inkonsistenz, die sich in den empirischen Daten widerspiegelt, wird häufig auf einen Mangel in der theoretischen Fundierung des Konstrukts zurückgeführt (Praetorius & Gräsel, 2021). Um das Konstruktverständnis zu schärfen, wurden jüngst Überlegungen artikuliert, kognitive Aktivierung fachlich zu spezifizieren (Begrich et al., 2023) oder die bisherigen Basisdimensionen von Unterrichtsqualität weiterhin als generisch zu interpretieren und um weitere, fachspezifische Basisdimensionen zu ergänzen (z. B. Lipowsky & Bleck, 2019).

Guttke, J. & Porsch, R. (2024). Inhaltsvalidierung von Fragebogenitems zu kognitiver Aktivierung im Englischunterricht der Grundschule mittels Expert:innenbefragung. In V. Lohe, A. Lindl & P. Kirchhoff (Hrsg.), *Unterrichtsqualität in schulischen Fremdsprachen* (S. 57–84). https://doi.org/10.31244/9783830999201.03

In der Fremdsprachenforschung wurde die Frage nach einer fachspezifischen Konstruktdefinition und -operationalisierung kognitiver Aktivierung bisher nicht systematisch untersucht. Dies ist jedoch eine notwendige Voraussetzung für die valide Erfassung des Konstrukts. Für das Fach Englisch existieren zwar erste Beobachtungsinstrumente (z. B. Kersten et al., 2018) und Fragebogenskalen (z. B. Porsch & Wilden, 2022), die kognitive Aktivierung im Fach Englisch zu erfassen vermögen (Lindl et al., in diesem Band). Der Entwicklungsprozess für diese Instrumente und die Prüfung von Konstruktvalidität als ein Schritt zur qualitativen Eignung wurden bislang jedoch nicht beschrieben. Aufgrund der konzeptionellen Unschärfe kognitiver Aktivierung ist insbesondere die Diskussion über Inhaltsvalidität von zentraler Bedeutung. Dies gilt umso mehr, wenn die mit diesen Instrumenten erhobenen Daten Einsatz in der evidenzbasierten Unterrichtsentwicklung finden sollen (Grünkorn et al., 2018).

Der vorliegende Beitrag adressiert das skizzierte Forschungsdesiderat in dreifacher Hinsicht. Erstens werden Ergebnisse einer Online-Expert:innenbefragung zur Inhaltsvalidität neukonstruierter Items zur Erfassung kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 mittels Schüler:innen- und Lehrkräftefragebögen dargelegt. Zweitens werden anhand der Ergebnisse Implikationen hinsichtlich der Konstruktoperationalisierung und Itemrevision abgeleitet. Schließlich leistet die Studie einen Beitrag zur Fragebogenentwicklung in der Fremdsprachenforschung, indem sie die Online-Expert:innenbefragung als eine Methode zur Entwicklung von Items aufzeigt, die bislang selten zur Anwendung gekommen ist.

2 Forschungsstand

2.1 Kognitive Aktivierung und die Notwendigkeit einer fachlichen Konstruktspezifikation

Dieser Abschnitt fasst jüngste Entwicklungen in der Erforschung kognitiver Aktivierung als Basisdimension von Unterrichtsqualität zusammen. Dazu werden zunächst verschiedene Facetten kognitiver Aktivierung begrifflich voneinander getrennt, die im bisherigen Forschungsdiskurs nur implizit unterschieden werden. Daraufhin werden Potenziale einer fachlichen Spezifikation kognitiver Aktivierung genannt, die in einer Konstruktdefinition für den Englischunterricht der Primarstufe münden.

Ursprünglich stellte kognitive Aktivierung im Kontext der TIMS-Studie einen faktorenanalytisch identifizierten Oberbegriff dar, der Unterrichtsmerkmale zur „Komplexität von Aufgabenstellungen und Argumentationen und die Intensität des fachlichen Lernens“ (Klieme et al., 2001, S. 51) umfasste. Seit dieser Veröffentlichung wurden die drei Basisdimensionen von Unterrichtsqualität über den Mathematikunterricht hinaus intensiv rezipiert und empirisch untersucht. Dies führte gleichzeitig zu einer Diffusion des Konstrukts kognitiver Aktivierung aufgrund

variierender theoretischer Verankerungen und Operationalisierungen (Schreyer et al., 2022). Derzeit erscheint ein Verständnis von kognitiver Aktivierung als Anregung „zum vertieften Nachdenken und zu einer elaborierten Auseinandersetzung mit dem Unterrichtsgegenstand" (Lipowsky, 2020, S. 92) konsensfähig.

Die Unbestimmtheit der Konstruktdefinition stellt eine Herausforderung in der Erforschung des Gegenstands dar, da die Grenzen zu weiteren Dimensionen der Unterrichtsqualität verschwimmen. Dies hat zur Folge, so stellen Lipowsky und Hess (2019) fest, dass unter kognitiver Aktivierung „auch inhaltlich teilweise recht unterschiedliche Maßnahmen der Lehrperson und/oder Aktivitäten der Lernenden verstanden werden" (S. 81). In einer Zusammenschau ausgewählter empirischer Studien identifizieren sie zehn Facetten kognitiver Aktivierung, die von der ko-konstruktiven Weiterentwicklung von Schüler:innenvorstellungen im Unterrichtsgespräch bis hin zur Adaptivität des Unterrichts reichen. Daraus leiten sie zwölf fachunabhängige Maßnahmen zur kognitiven Aktivierung der Lernenden für die Unterrichtspraxis ab. Zu diesen Maßnahmen zählen unter anderem die „[s]ystematische Variation von Aufgaben(teilen) und Anregung zur Entdeckung von Regelmäßigkeiten" sowie der Einbezug von Aufgaben, „die über die Anwendung von Routinen hinausgehen" (Lipowsky & Hess, 2019, S. 92).

Das Syntheseframework zu Unterrichtsqualität (Praetorius et al., 2020) ergänzt die drei Basisdimensionen auf Grundlage einer Synthese von zwölf internationalen *frameworks* zur Erfassung von Unterrichtsqualität um vier weitere Dimensionen. Darin wird – trotz häufiger Verweise auf die Abhängigkeit von kognitiver Aktivierung und Lerngegenständen – eine Trennung zwischen kognitiver Aktivierung und Unterrichtsinhalten vorgenommen, indem sie als zwei separate Unterrichtsqualitätsdimensionen formuliert werden. Kognitiver Aktivierung werden vier Subdimensionen zugeschrieben: (1) Auswahl und (2) Einsatz fachlich gehaltvoller und auf das kognitive Niveau der Schüler:innen abgestimmter Aufgaben sowie (3) Unterstützung der kognitiven Aktivität und (4) des metakognitiven Lernens der Schüler:innen anhand kognitiv aktivierender Aufgaben.

Bei genauerer Betrachtung dieser Konstruktbeschreibungen lassen sich aus den unterschiedlichen Indikatoren kognitiver Aktivierung drei separate Facetten herausarbeiten (Kleickmann et al., 2020; Rieser & Decristan, 2023): (1) das Potenzial zu kognitiver Aktivierung, (2) die kognitive Unterstützung und (3) die kognitive Aktiviertheit. Als Potenziale zu kognitiver Aktivierung gelten grundsätzlich alle Unterrichtsmerkmale, von denen begründet angenommen wird, dass sie der kognitiven Aktivierung von Schüler:innen dienlich sind. Inwiefern diese Potenziale aufseiten der Schüler:innen tatsächlich in kognitiver Aktiviertheit resultieren, ist von interindividuellen Faktoren und der Wahrnehmung des Unterrichtsangebots abhängig. Maßnahmen der kognitiven Unterstützung zielen darauf ab, die Ansprüche komplexer Lernumgebungen – die gewissermaßen notwendig für kognitive Aktivierung sind – zu reduzieren, sodass eine Bearbeitung durch unterschiedlich leistungsstarke Schüler:innen möglich wird. Ob diese drei Facetten Teil eines gemeinsamen Konstrukts kognitiver Aktivierung sind oder in Teilen separate Basisdimensionen der Unterrichtsqualität bilden, ist noch unklar.

Unstrittig ist hingegen, dass die drei Facetten kognitiver Aktivierung konform mit dem Angebot-Nutzungs-Modell schulischer Lehr-Lernprozesse sind (Helmke, 2014).

Die stark divergierenden Definitionen und Operationalisierungen kognitiver Aktivierung lassen sich zum Teil auf die drei beschriebenen Facetten zurückführen. Während die von Lipowsky und Hess (2019) identifizierte Offenheit von Problemstellungen ein Potenzial zu kognitiver Aktivierung darstellt, bildet die Individualisierung von Unterricht einen Indikator kognitiver Unterstützung. Die Unterteilung in drei Facetten kognitiver Aktivierung birgt Implikationen nicht nur auf inhaltlicher, sondern auch auf methodischer Ebene. Rieser und Decristan (2023) zeigen beispielsweise, dass sich Skalen, die konzeptionell zwischen dem Potenzial zu kognitiver Aktivierung und kognitiver Aktiviertheit unterscheiden, auch empirisch auf Individual- und Klassenebene voneinander trennen lassen und auf Klassenebene nicht signifikant miteinander korrelieren. Vor diesem Hintergrund ist es ratsam, bei der Neuentwicklung von Instrumenten zur Erfassung von kognitiver Aktivierung bereits während der Itemkonstruktion die drei Facetten des Konstrukts zu berücksichtigen.

Das Potenzial zu kognitiver Aktivierung wurde bisher größtenteils in Unterrichtsmaterialien wie Lehrwerken (für das Fach Englisch: Wilden et al., eingereicht) sowie Lernaufgaben verortet und generisch untersucht, jedoch nur selten als Teil des Lehrkräftehandelns begriffen. Maier et al. (2010) schlagen beispielsweise ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben vor. Dazu beschreiben sie sieben Dimensionen zur Kategorisierung von Aufgaben: Art des Wissens, die zur Bewältigung der Aufgabe notwendigen kognitiven Prozesse, Anzahl der bei der Aufgabenbearbeitung aktivierten Wissenseinheiten, Offenheit der Aufgabenstellung, Lebensweltbezug, sprachlogische Komplexität und Repräsentationsformen des Wissens. Wird das Kategoriensystem zur Analyse von Aufgabenmaterial (*task as workplan*) eingesetzt, lässt sich daran eine Aussage über das Potenzial zur kognitiven Aktivierung treffen, das wiederum von der Implementation des Aufgabenmaterials in das Unterrichtsgeschehen (*task in process*) und den dadurch initiierten Lernprozessen zu unterscheiden ist.

Die beispielhaft aufgeführten – und bei Weitem nicht vollständigen – Facetten und Indikatoren kognitiver Aktivierung tragen dazu bei, das Konstrukt im Sinne einer extensionalen Nominaldefinition einzugrenzen (Döring & Bortz, 2016). In der Anwendung und Operationalisierung erfordern diese Indikatoren in Abhängigkeit von dem jeweils betrachteten Schulfach dennoch einen Transferschritt, zu dessen Durchführung fachdidaktische Forschung unerlässlich ist. So stellt sich beispielsweise die Frage, welche Charakteristika im Englischunterricht fachlich gehaltvolle Aufgaben kennzeichnen oder welche Bestandteile einer Aufgabe auf ihre sprachlogische Komplexität hin zu prüfen sind (lediglich die Aufgabeninstruktion oder auch der Ausgangstext bei einer Aufgabe zum Leseverstehen; Wilden et al., eingereicht). Des Weiteren werden einige dieser Fragen erst durch

einen komparativen Ansatz aus Perspektive unterschiedlicher Fachdidaktiken beantwortbar (z. B. Lindl et al., in diesem Band; Praetorius et al., 2020).

Im Forschungsdiskurs zu Unterrichtsqualität bildet sich die Tendenz zu einem fachspezifischen Verständnis kognitiver Aktivierung in zweierlei Hinsicht ab. Einige Forscher:innen sprechen sich für eine „Erweiterung kognitiver Aktivierung über Fächer hinweg oder aber lediglich für einzelne Fächer hin zu einer kognitiv-motorischen, kognitiv-ästhetischen oder kommunikativ-kognitiven Aktivierung" (Praetorius & Gräsel, 2021, S. 178) aus. Andere Ansätze schlagen hingegen vor, „diese ergänzenden Aspekte nicht als Teil kognitiver Aktivierung, sondern als eigenständige Dimensionen zu behandeln" (Praetorius & Gräsel, 2021, S. 178; Lindl et al., 2024). Ungeachtet dieser Entscheidung birgt die fachliche Spezifikation kognitiver Aktivierung das Potenzial, die Lernwirksamkeit des Fachunterrichts differenzierter und damit den Gegenstand angemessener zu untersuchen.

Für die fachspezifische Operationalisierung kognitiver Aktivierung sind die Bildungsziele eines Fachs von zentraler Bedeutung. In den Bildungsstandards für den ersten und mittleren Schulabschluss wird die Entwicklung funktionaler kommunikativer Kompetenz als übergeordnetes Ziel des Fremdsprachenunterrichts postuliert (KMK, 2023). Ergänzend werden Kompetenzbereiche wie interkulturelle Kompetenz, Sprachbewusstheit oder Text- und Medienkompetenz als nebengeordnete Ziele benannt. Aufgrund ihrer nebengeordneten Rolle und um den Umfang der zu entwickelnden Fragebögen zu begrenzen, werden diese Kompetenzbereiche in diesem Forschungsprojekt nicht explizit adressiert, obgleich sie vereinzelt in Items abgebildet sind. Dies soll nicht bedeuten, dass den fünf nebengeordneten Kompetenzbereichen keine Relevanz hinsichtlich der Qualität des Englischunterrichts zugeschrieben wird; stattdessen steht die Untersuchung kognitiver Aktivierung während des Zweitspracherwerbs im Zentrum des Forschungsinteresses.

Zum Fremdsprachenunterricht liegen aus dem deutschsprachigen Raum bisher nur vereinzelt konzeptionelle Überlegungen und empirische Befunde vor, die sich im weitesten Sinne der Untersuchung von kognitiver Aktivierung zuordnen lassen. Guttke (2023) identifiziert in einem systematischen Forschungsreview sechs Aspekte, die kognitiver Aktivierung im Fremdsprachenunterricht dienlich seien: die Auseinandersetzung mit einer kommunikativen Problemstellung, der zielgerichtete Einsatz der Zielsprache in einem bedeutsamen Kontext, die kritische Reflexion des eigenen Sprachhandelns, vielfältiger und verständlicher *language input*, Gelegenheit zur Produktion von *language output* während der Interaktion in der Zielsprache sowie begleitendes *corrective feedback*. Anhand dieser Überlegungen lässt sich eine vorläufige, fachspezifische Konstruktdefinition formulieren: Kognitive Aktivierung im Fremdsprachenunterricht entspricht einer möglichst intensiven Verarbeitung der Zielsprache. Unter Berücksichtigung der Forschung zu *(Instructed) Second Language Acquisition* (Gass et al., 2020, S. 575–589; Loewen, 2020) und der sechs Aspekte kognitiver Aktivierung im Fremdsprachenunterricht (Guttke, 2023) drückt sich eine solche möglichst intensive Verarbeitung der Zielsprache potenziell in drei Prozessen aus: (1) in der Transformation von *language*

input zu *intake*, (2) in Schüler:innenreaktionen auf *corrective feedback* (*uptake*) sowie (3) während der Produktion zielsprachlicher Äußerungen (*comprehensible/ pushed output*).

2.2 Erfassung kognitiver Aktivierung mittels Fragebögen

Im Folgenden werden schriftliche Befragungen von Schüler:innen und Lehrkräften mittels Fragebögen als Methode zur Erfassung von Unterrichtsqualität einschließlich kognitiver Aktivierung sowie empirische Befunde zu ihrer psychometrischen Güte präsentiert. Anhand dieser Befunde wird die Relevanz von Itemperspektive und -referenz in der Itemformulierung für die Entwicklung valider Fragebögen diskutiert, woraus Implikationen für die Itemkonstruktion im nachfolgend beschriebenen Promotionsprojekt abgeleitet werden.

Neben der Analyse von Aufgaben hinsichtlich ihres Potenzials zu kognitiver Aktivierung mittels Kodierschemata stellen Erhebungen mit Fragebögen, die eine Anzahl inhaltlich zusammengehöriger Items bzw. Aussagen zur Bewertung eines Sachverhalts umfassen, ein üblicherweise genutztes und ökonomisches Instrumentarium dar, um Unterrichtsqualität einschließlich kognitiver Aktivierung quantitativ zu messen. Die aus den Fragebogenerhebungen gewonnenen Kennwerte stellen Indikatoren für eine hohe oder geringe Qualität dar und können wiederum mit weiteren Indikatoren von Unterrichtsqualität, Prozessmerkmalen des Unterrichts, Merkmalen der Lehrkraft oder der Schüler:innen sowie *outcomes* des Unterrichts wie Schüler:innenleistungen in einen Zusammenhang gebracht werden (vgl. Lindl et al., in diesem Band). Theoretische Zusammenhänge dieser verschiedenen Merkmalsgruppen lassen sich mithilfe des Angebot-Nutzungs-Modells des Unterrichts von Helmke (2014) erklären. Zentral sind in dem Modell neben der Annahme von Zusammenhängen die Unterscheidung von Merkmalen der Lehrperson, des Unterrichts als ein Angebot, den Lernaktivitäten, d. h. die Nutzung des Angebots seitens der Schüler:innen, und den Wirkungen bzw. Ertrag wie den Leistungen oder Veränderungen non-kognitiver Merkmale (z. B. Selbstwirksamkeitsüberzeugungen). Dahingehend stellt sich für die Entwicklung von Skalen zur Erfassung von kognitiver Aktivierung im Unterricht die Frage, ob (a) das *Potenzial* zu kognitiver Aktivierung oder (b) die *Nutzung* des unterrichtlichen Angebots erfasst werden sollen (s. Abschn. 2.1). Nachteil ist für letztgenannten Ansatz, dass damit Aspekte kognitiver Aktiviertheit abgebildet werden, die vorrangig für Schüler:innen zugänglich sind und nur eingeschränkt durch die Berücksichtigung weiterer Perspektiven auf ihre Gültigkeit geprüft werden können. Vor diesem Hintergrund fokussieren die im vorliegenden Beitrag präsentierten Items das Potenzial zu kognitiver Aktivierung.

Weiterführend lassen sich zur Einschätzung der Qualität des Unterrichts in Bezug auf das Potenzial zur kognitiven Aktivierung der Schüler:innen drei Perspektiven unterscheiden: die der Lehrkraft, die der Schüler:innen (individuell und in Form von aggregierten Klassenmerkmalen) und die von externen Beob-

achter:innen (einschließlich Videographie). Alle Perspektiven besitzen bestimmte Potenziale als auch Nachteile (z. B. Gruehn, 2000) und empirische Arbeiten zeigen, dass sich die Wahrnehmung der verschiedenen Gruppen unterscheidet (z. B. Fauth et al., 2014). Göllner et al. (2016) diskutieren insbesondere die Erfassung von Unterrichtsqualität aus Schüler:innensicht. Die Perspektive der Schüler:innen hat unter anderem den Vorteil, dass eine hohe Anzahl an Beurteiler:innen zur Verfügung steht, um das Unterrichtsgeschehen angemessen abzubilden. Laut den Autor:innen liegen relativ wenige Befunde zur psychometrischen Güte von Schüler:innenurteilen vor; die verfügbaren Befunde deuten jedoch darauf hin, dass sich Schüler:innenurteile als Datenquelle mit hinreichender Validität zur Erfassung von Unterrichtsqualität eignen. Insgesamt seien jedoch alle Perspektiven relevant und keine zu bevorzugen, da sich jede Perspektive für bestimmte Konfigurationen von Beurteilungsgegenstand und Itemreferenz besonders gut eignet (S. 70).

Schließlich können in Fragebögen neben den verschiedenen Personen Bewertungen von Merkmalen des Unterrichts oder eine Bewertung der Unterrichtsqualität in Bezug auf (a) eine spezifische Unterrichtsstunde oder (b) retrospektiv summierend Unterricht (z. B. bezogen auf das zurückliegende Schuljahr) in einem Fach eingeschätzt werden. In Bezug auf (a) kann es sich um spezifische, einfach beobachtbare Geschehnisse im Unterricht handeln. Alternativ sind hoch-inferente Beobachtungen (im Sinne von Ratings) möglich, die von der bzw. dem Beurteiler:in verlangen, „über das konkret beobachtete Verhalten hinaus auf abstrakte Sachverhalte oder allgemeine Verhaltenstendenzen zu schließen“ (Göllner et al., 2016, S. 64), was zum Erreichen reliabler Ergebnisse bestenfalls mit einer vorangegangenen Schulung verbunden ist. Grundsätzlich verlangt die Bewertung einer Aussage bzw. die Beantwortung eines Items unterschiedlich komplexe kognitive Prozesse.

Neben grundsätzlichen Anforderungen an die Itemkonstruktion (z. B. Eindeutigkeit, sprachliche Verständlichkeit), um die Beantwortung zu ermöglichen, können laut Göllner et al. (2016, S. 70) weitere Anforderungsmerkmale unterschieden werden: (a) Adressat:innenbezug des Items (Wer wird von den Items adressiert bzw. wer soll bewertet werden? Gesamte Gruppe/Klasse oder einzelne:r Schüler:in), (b) die Wahrnehmungsperspektive (Aus welcher Sicht erfolgt die Beurteilung? gesamte Gruppe/Klasse oder einzelne/r Schüler:in) und (c) der Zeitbezug (Auf welchen Zeitraum bezieht sich die Beurteilung eines Items?), ein Aspekt, der bereits oben thematisiert wurde. Ergänzt um die Interaktion zwischen Lehrer:innen und Schüler:innen entwickelten Fauth et al. (2020) eine Matrix zur Item-Referenz (s. Abb. 1).

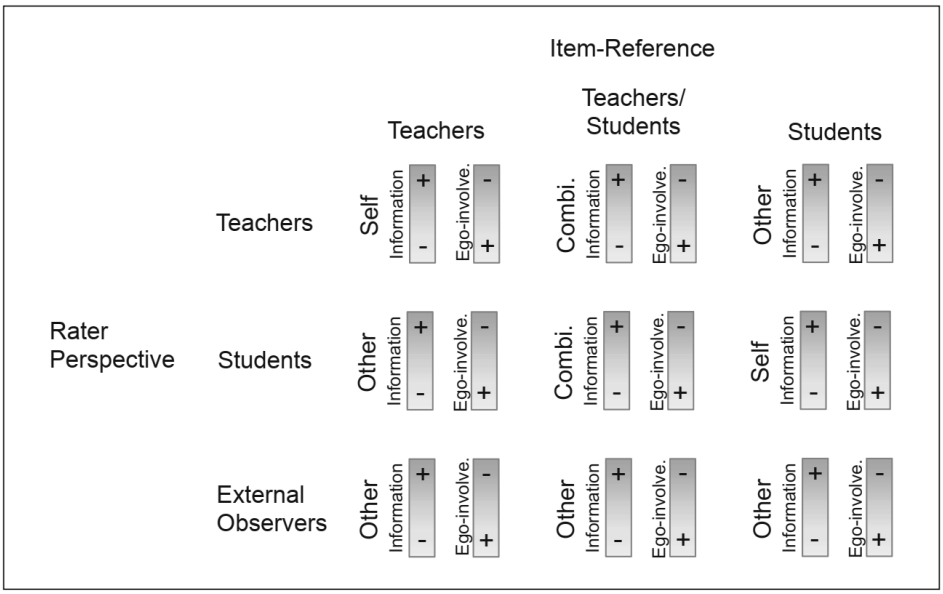


Abbildung 1: Reference perspective matrix (Fauth et al., 2020, S. 144)

Die Matrix zeigt einerseits die drei Gruppen, die eine Beantwortung vornehmen können (Lehrkräfte, Schüler:innen, externe Beobachter:innen). Sie unterscheidet anderseits darin, dass sich der Inhalt eines Items auf das Handeln der Lehrkraft, die Schüler:innen oder die Interaktion zwischen beiden Gruppen beziehen kann. Schließlich wird klassifiziert, ob sie ihr eigenes Verhalten oder das anderer bewerten sollen. Die kritische Frage stellt sich, inwieweit jeweils für letzteren Punkt stets ausreichend Informationen zur Verfügung stehen (bspw. um als Schüler:in eine Aussage für die gesamte Klasse treffen zu können). Für die Bewertung des eigenen Verhaltens ist anzunehmen, dass diese bestimmten Verzerrungen unterworfen sein kann (u. a. Schutz des Selbstwerts oder im Sinne sozialer Erwünschtheit). Göllner et al. (2016) und Fauth et al. (2020) zeigen auf, dass Inventare zur Messung von Unterrichtsqualität die verschiedenen Anforderungen abbilden und mitunter sogar unterschiedliche Perspektiven innerhalb einer Skala eingenommen werden sollen. Auch der Zeitbezug ist nicht immer eindeutig für die Beurteiler:innen erkennbar.

Um kognitive Aktivierung angemessen zu operationalisieren, bedarf es daher neben der Festlegung auf eine fachspezifische Konstruktdefinition eines systematischen Vorgehens bei der Formulierung der Items und die Sicherstellung, dass die Befragten die Items mit ihnen zur Verfügung stehenden Informationen eindeutig für einen festgelegten Zeitraum beantworten können. In der nachfolgend vorgestellten Studie wurden diese Voraussetzungen in der Itemkonstruktion wie folgt umgesetzt: Bei den Lehrkräfteitems wurde darauf geachtet, dass die Items in der Ich-Perspektive formuliert sind und sich die Iteminhalte (sofern möglich) ausschließlich auf das Lehrkräfteverhalten beziehen. Die Schüler:innenitems wurden so formuliert, dass sie Beurteilungen des Lehrkräfteverhaltens auf Individual-

statt auf Klassenebene elizitieren. In den Items beider Zielgruppen besteht eine klare Referenz hinsichtlich des zeitlichen (zurückliegende Englischstunde) und inhaltlichen Bezugs zu einem möglichst beobachtbaren Verhalten.

2.3 Expert:innenbefragungen als Instrument zur Inhaltsvalidierung von Fragebögen

Da der vorliegende Beitrag eine Validierungsstudie darstellt, fasst dieser Abschnitt das Konzept der Inhaltsvalidität zusammen. Darüber hinaus werden anhand empirischer Studien aus der Fachdidaktik und den Bildungswissenschaften methodische Zugänge zum Erreichen von Inhaltsvalidität aufgezeigt, welche die Grundlage für das in diesem Beitrag gewählte methodische Vorgehen bilden.

Im Zuge der Test- und Fragebogenentwicklung bildet die Überprüfung der Güte eines Instruments einen zentralen Bestandteil des Forschungsprozesses. Ursprünglich wurde Validität als ein einem Fragebogen inhärentes Merkmal begriffen. Inzwischen wird das Ausmaß des Gütekriteriums hingegen anhand der Interpretation und weiteren Verwendung von *scores*, die aus der Anwendung eines Fragebogens resultieren, bewertet (Messick, 1995, S. 742). Damit wird die Validierung eines Fragebogens zu einem fortdauernden und argumentativen Prozess, für den es je nach Anwendungskontext geeignete Evidenz anzuführen gilt.

Trotz dieses zunehmend integrativen Validitätsbegriffs kann es zur Identifikation geeigneter Evidenz nützlich sein, folgende Validitätsfacetten voneinander zu unterscheiden: Kriteriumsvalidität, Konstruktvalidität sowie schließlich Inhaltsvalidität, die Gegenstand des vorliegenden Beitrags ist (Moosbrugger & Kelava, 2020). Anhand der Inhaltsvalidität wird untersucht, inwiefern „die Items eines Tests/Fragebogens eine repräsentative Stichprobe an Verhaltens- und Erlebensweisen aus jenem Itemuniversum (d. h. allen merkmalsrelevanten Verhaltens- und Erlebensweisen) darstellen, mit dem das interessierende Merkmal vollständig erfasst werden könnte“ (Moosbrugger & Kelava, 2020, S. 32). Aus dieser Definition lassen sich mit der Repräsentanz und inhaltlichen Relevanz von Items zwei Kriterien zur Überprüfung von Inhaltsvalidität formulieren (Klauer, 1984; Messick, 1995). Enthält ein Fragebogen beispielsweise zu wenige Items, um ein komplexes Konstrukt zu erfassen, resultiert dies in „*construct underrepresentation*“ (Messick, 1995, S. 742). Andererseits kann ein Fragebogen zwar ausreichend Items enthalten, von denen einige jedoch nicht inhaltlich relevant für das zu erfassende Konstrukt sind und im ungünstigsten Fall sogar ein anderes Konstrukt erfassen („*construct-irrelevant variance*“; Messick, 1995, S. 742). Somit lässt sich Inhaltsvalidität sowohl auf Itemebene als auch holistisch auf Fragebogenebene betrachten.

Trotz der aufgezeigten Relevanz findet eine explizite Untersuchung von Inhaltsvalidität in fremdsprachendidaktischen Forschungsarbeiten bislang selten statt. Grundlage für die Überprüfung von Inhaltsvalidität bildet die Formulierung einer Konstruktspezifikation, anhand derer alle konstruierten Items unter Verweis auf Bezugstheorien des Konstrukts legitimiert werden können (Döring & Bortz,

2016). Davon ausgehend lassen sich Annahmen zur systematischen Konstruktion eines Itempools innerhalb eines nomologischen Netzes generieren, um ein ausreichendes Maß an Repräsentanz sicherzustellen (Klauer, 1984). Die inhaltliche Relevanz des Itempools lässt sich weitaus weniger systematisch prüfen, da die Inhaltsvalidierung „theoretisch-argumentativ und gestützt durch Urteile von Fachexperten" (Döring & Bortz, 2016, S. 446) erfolgt, statt sie anhand eines numerischen Kennwertes zu quantifizieren. Wie eine solche Expert:innenbeurteilung methodisch umgesetzt wird, soll nachfolgend anhand ausgewählter Studien illustriert werden. Aufgrund der seltenen Diskussion von Inhaltsvalidität in fremdsprachendidaktischer Forschung wird dazu auch auf Studien anderer Fachdidaktiken sowie der empirischen Bildungsforschung zurückgegriffen.

Eine in der fachdidaktischen Forschung relativ etablierte Methode zur Erfassung von Expert:innenurteilen bilden Delphi-Befragungen. Dabei beantworten Expert:innen schriftlich über mehrere Runden hinweg Items, deren Ergebnisse im Anschluss an jede Runde anonymisiert und kumuliert präsentiert werden. Dadurch wird eine schrittweise, moderierte Konsensbildung unter den beteiligten Expert:innen möglich. Weber et al. (2020) führten beispielsweise eine drei Runden umfassende und über ein Jahr andauernde Delphi-Befragung zur Konstruktion von Leitlinien für die Konzeption einer Lehrkräftefortbildung zum Thema Quantenphysik durch. Da es wünschenswert ist, dass zumindest ein Teil der Expert:innen über mehrere Runden an Delphi-Befragungen teilnehmen, verdeutlicht diese Studie bereits, welche hohen Ansprüche die Methode an die zeitlichen Ressourcen ihrer Teilnehmer:innen stellt.

Im Rahmen des interdisziplinären Projekts FALKO beschreibt Kirchhoff (2016, 2017) die Entwicklung eines Testinstruments zur Erfassung fachspezifischen professionellen Wissens von Englischlehrkräften. Zur Prüfung der Inhalts- und Augenscheinvalidität wurden 15 Expert:inneninterviews – darunter elf Englischlehrkräfte unterschiedlicher Schulformen der Sekundarstufe und vier Fachdidaktiker:innen – im Modus des lauten Denkens durchgeführt. Zu den Inhalten oder der Auswertung dieser Interviews werden in den gesichteten Publikationen keine näheren Angaben gemacht. Zudem wurden in der Vorstudie 125 Lehramtsstudierende mit dem Fach Englisch sowie Englischlehrkräfte schriftlich zur Relevanz der Testitems für die Unterrichtspraxis befragt („Ich halte diesen Inhalt für die Ausbildung von Lehrkräften für nicht relevant/kaum relevant/ziemlich relevant/sehr relevant"; Kirchhoff, 2016, S. 88).

Schließlich werden in einer Vielzahl an Forschungsarbeiten (z. B. Hofrichter et al., in diesem Band; Jenßen et al., 2015; Senkbeil et al., 2013; Wibowo & Heemsoth, 2019) Varianten schriftlicher Expert:innenbefragungen zur Prüfung der Inhaltsvalidität eingesetzt. Dabei werden anhand zuvor explizit formulierter Selektionskriterien identifizierte Expert:innen mittels Ratingskalen zur Beurteilung von Items aufgefordert. Von Interesse sind dabei beispielsweise die Relevanz und Repräsentanz der Items für das zu messende Konstrukt oder die Zugehörigkeit der Items zu unterschiedlichen Wissensfacetten in Professionstests. Vereinzelt werden die Ratingskalen um Freitextantworten ergänzt, um schriftliche Kommentare der

Expert:innen zu ermöglichen. Die Stichprobengröße variiert in den vier gesichteten Publikationen zwischen sechs und 55 Teilnehmer:innen. In gewisser Hinsicht stellt die schriftliche Expert:innenbefragung damit eine Mischform der beiden zuvor beschriebenen Zugänge dar, indem sie durch ihren quantitativen Schwerpunkt einerseits die wenig zeitintensive Befragung möglichst vieler Expert:innen erlaubt und andererseits Begründungen der Expert:innen zu ihren Urteilen einfordert.

3 Forschungsfragen

Dieser Beitrag ist Teil eines Promotionsprojekts, in dessen Rahmen Schüler:innen- und Lehrkräftefragebögen zur Erfassung des Potenzials zu kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 entwickelt, erprobt und validiert werden. Damit verfolgt der Beitrag eine der einleitend formulierten Forschungsaufgaben in Bezug auf fremdsprachliche Unterrichtsqualität (Lindl et al., in diesem Band). Dazu wurde in einem ersten Schritt der Forschungsstand zu kognitiver Aktivierung im Fremdsprachenunterricht im deutschsprachigen Raum systematisch aufgearbeitet (Guttke, 2023). Auf Grundlage dieser Ergebnisse und Befunde der *Instructed Second Language Acqusition* (Loewen, 2020) wurde eine vorläufige Konstruktdefinition formuliert (vgl. Abschn. 2.1). Diese diente der Konstruktion von Items, die daraufhin in mehreren Teilstudien validiert wurden. Neben der Durchführung kognitiver Interviews zur Prüfung der Iteminterpretation und einer Pilotierung zur deskriptiv-statistischen Itemanalyse wurde die nachfolgend vorgestellte Online-Expert:innenbefragung zur Inhaltsvalidierung der Items durchgeführt. Mithilfe der Daten dieser Befragung sollen die folgenden Forschungsfragen beantwortet werden:

1. In welchem Ausmaß bilden die theoriebasiert konstruierten Fragebogenitems zur Erfassung kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 laut Expert:innenurteil relevante Aspekte hinsichtlich des Konstrukts ab?
2. Welche Itemmerkmale bewerten die Expert:innen positiv/negativ hinsichtlich der Itemkonstruktion (z. B. Itemlänge oder Kongruenz von Schüler:innen- und Lehrkräfteversion eines Items)?

Das weiterführende Ziel der Beantwortung von Forschungsfrage 1 soll sein, ambige oder unpassende Items zu identifizieren, die laut Expert:innenmeinung entweder keinen Aspekt kognitiver Aktivierung abbilden oder möglicherweise solche Aspekte erfassen, die einer anderen Dimension von Unterrichtsqualität zugehörig sind. Gleichzeitig können damit Leerstellen aufgezeigt und neue Items konstruiert werden, um Aspekte kognitiver Aktivierung zu erfassen, die im Itempool bislang nicht berücksichtigt wurden. Forschungsfrage 2 bietet das Potenzial, die Itemkonstruktion durch die Expertise der Teilnehmer:innen an der Online-Befragung evaluieren zu lassen und daraus Hinweise für die Revision der Items abzuleiten.

4 Methodisches Vorgehen

4.1 Rekrutierung von Expert:innen und Stichprobe

Die Eignung zur Inhaltsvalidierung des Itempools setzt seitens der Teilnehmer:innen Expertise hinsichtlich kognitiver Aktivierung voraus. Diese Expertise kann, so wurde im Rahmen der Studie angenommen, entweder aus der wissenschaftlichen Auseinandersetzung mit dem Forschungsgegenstand oder der Reflexion unterrichtspraktischer Erfahrung erwachsen. Folglich ergaben sich drei Personengruppen für die Teilnehmer:innenakquise: Hochschullehrer:innen, Grundschullehrkräfte und Fachleitungen der Zentren für schulpraktische Lehrerausbildung (ZfsL) in Nordrhein-Westfalen (NRW). Zur Berücksichtigung bei der Akquise sollten Hochschullehrer:innen forschend in der Fremdsprachendidaktik oder der empirischen Bildungsforschung tätig sein und in den zurückliegenden fünf Jahren mindestens eine Publikation zu Unterrichtsqualität oder kognitiver Aktivierung veröffentlicht haben. Bei den Grundschullehrkräften wurde ein Abschluss des Vorbereitungsdienstes vorausgesetzt. An den ZfsL wurden ausschließlich Fachleitungen für das Fach Englisch in der Grundschule kontaktiert.

Die Akquise der Teilnehmer:innen für die Online-Expert:innenbefragung erfolgte per E-Mail. In einer ersten E-Mail wurde die Teilnahmebereitschaft der gemäß den Auswahlkriterien identifizierten Expert:innen erfragt. Diese Einladung enthielt nähere Informationen zum Forschungsprojekt sowie zum Umfang und Ablauf der Online-Befragung. Alle Expert:innen, die dieser Einladung folgten, erhielten mit einer zweiten E-Mail einen personalisierten Zugangslink zur Online-Befragung und eine schriftliche Konstruktdefinition kognitiver Aktivierung im Englischunterricht der Primarstufe, um eine gemeinsame Ausgangsbasis für das Rating der Items unter den Expert:innen zu gewährleisten.

Letztendlich nahmen 28 Expert:innen an der Online-Befragung teil, von denen 24 die Online-Befragung vollständig absolvierten. Die vier unvollständigen Datensätze wurden von der weiteren Datenauswertung ausgeschlossen. Die Expert:innen setzen sich aus zwölf Hochschullehrer:innen und zwölf Grundschullehrkräften zusammen, von denen acht zusätzlich als Fachleitungen an ZfsL in NRW tätig sind. Durchschnittlich wiesen die Hochschullehrer:innen eine Berufserfahrung von elf Jahren (M = 10,92; SD = 3,57) und die Grundschullehrkräfte von zehn Jahren (M = 10,17; SD = 5,89) auf. Vier der Hochschullehrer:innen ordneten sich der empirischen Bildungsforschung, acht der Fremdsprachendidaktik zu. Hinsichtlich ihrer Qualifikation gaben sieben Grundschullehrkräfte Englisch als Fach in Lehramtsstudium und Vorbereitungsdienst an, wohingegen bei zwei Grundschullehrkräften das Fach Englisch Teil des Lehramtsstudiums, nicht jedoch des Vorbereitungsdienstes war. Drei Grundschullehrkräfte studierten das Fach Englisch nicht, sondern erwarben ihre Lehrqualifikation durch eine Nachqualifizierung.

4.2 Erhebungsinstrument

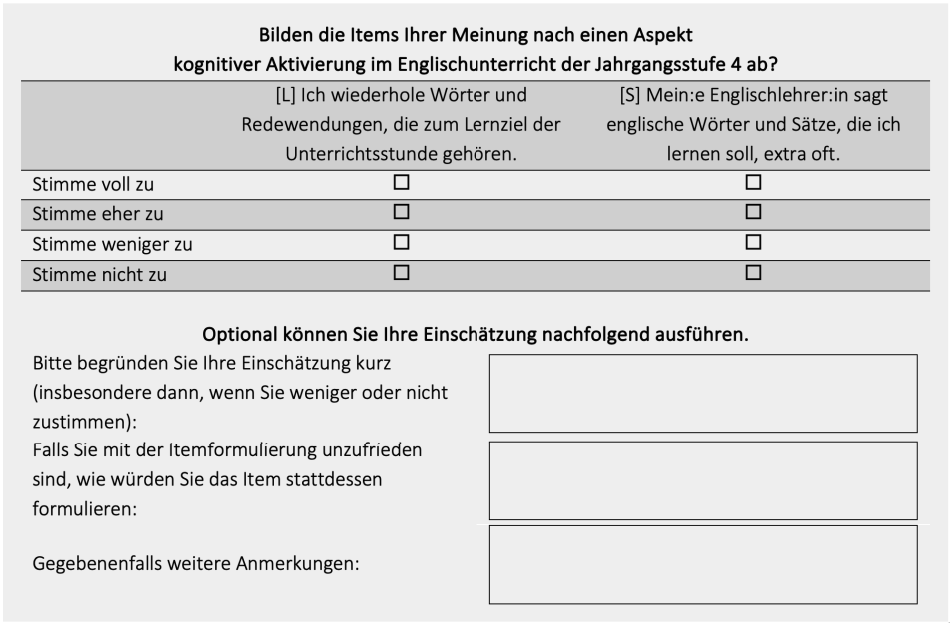
Die Online-Befragung fand im Zeitraum vom 04.05.–19.06.2022 statt und wurde über LimeSurvey durchgeführt. Als Erhebungsinstrument diente der gesamte theoriebasiert konstruierte Itempool, der zum damaligen Zeitpunkt 41 Itempaare verteilt auf vier Skalen (*language input*, *information processing*, *language output*, *corrective feedback*; vgl. Tab. 1) umfasste. Jedes Itempaar bestand aus einem Schüler:innen- und Lehrkräfteitem, welche dieselben Inhalte aus unterschiedlichen Perspektiven beschrieben. Die Lehrkräfteitems wurden in der Ich-Perspektive (Selbstbewertung) formuliert, wohingegen die Schüler:innenitems größtenteils Aussagen über die Englischlehrkraft enthielten (Fremdbewertung: „Mein:e Englischlehrer:in …").

Tabelle 1: Beispielitems zu den vier in der Online-Expert:innenbefragung bewerteten Skalen

Itemcode	Lehrkräfteitems	Schüler:innenitems
	Language input	
modifikation	Ich verändere meinen Sprachgebrauch im Englischen je nach Lernstand der Schüler:innen, mit denen ich spreche.	Wenn mein:e Englischlehrer:in mit mir auf Englisch spricht, verstehe ich das meiste, was sie/er sagt.
	Information processing	
erstsprache	Ich ermögliche den Schüler:innen, die englische Sprache mit ihrer Erstsprache zu vergleichen.	Im Englischunterricht vergleiche ich Englisch mit anderen Sprachen. Ich denke darüber nach, was bei den beiden Sprachen gleich und was verschieden ist.
	Language output	
nonverbal-aeussern	Wenn Schüler:innen ein englisches Wort nicht kennen, um einen Satz zu formulieren, dürfen sie sich non-verbal äußern (z. B. Gestik, Mimik, Symbolkarten).	Wenn ich ein Wort auf Englisch nicht weiß, darf ich Hilfsmittel benutzen (z. B. meine Hände, Karten mit Bildern).
	Corrective feedback	
inhaltsfehler	Ich korrigiere inhaltliche Fehler der Schüler:innen (z. B. „My mother is 9 years old.").	Mein:e Englischlehrer:in berichtigt mich, wenn ich etwas sage, das nicht stimmt (z. B. „My mother is 9 years old.").

Nach Abfrage einiger demografischer Merkmale wurde den Expert:innen nacheinander jedes Itempaar präsentiert (vgl. Abb. 2). Die Abfolge der Items innerhalb der Skalen wurde randomisiert. Die Repräsentanz und Relevanz eines Items konnte entlang einer vierstufigen Likert-Skala eingeschätzt werden. Zusätzlich standen den Expert:innen drei Freitextfelder für die Begründung ihrer Einschät-

zung, eines alternativen Formulierungsvorschlags sowie weitere Anmerkungen zur Verfügung.



Bilden die Items Ihrer Meinung nach einen Aspekt kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 ab?

	[L] Ich wiederhole Wörter und Redewendungen, die zum Lernziel der Unterrichtsstunde gehören.	[S] Mein:e Englischlehrer:in sagt englische Wörter und Sätze, die ich lernen soll, extra oft.
Stimme voll zu	☐	☐
Stimme eher zu	☐	☐
Stimme weniger zu	☐	☐
Stimme nicht zu	☐	☐

Optional können Sie Ihre Einschätzung nachfolgend ausführen.

Bitte begründen Sie Ihre Einschätzung kurz (insbesondere dann, wenn Sie weniger oder nicht zustimmen):

Falls Sie mit der Itemformulierung unzufrieden sind, wie würden Sie das Item stattdessen formulieren:

Gegebenenfalls weitere Anmerkungen:

Abbildung 2: Illustration des Frageformats aus der Online-Expert:innenbefragung anhand eines Beispielitems der Skala language input

4.3 Datenauswertung

Aufgrund der Verwendung eines geschlossenen und eines offenen Antwortformats resultierten aus der Online-Befragung sowohl quantitative als auch qualitative Daten, die auf unterschiedliche Weise ausgewertet wurden. Die Daten der vierstufigen Likert-Skalen zur Einschätzung der Items wurden zunächst dichotomisiert. Dazu wurden jeweils zwei Ausprägungen der Likert-Skala (*stimme nicht zu*/*stimme weniger zu* sowie *stimme voll zu*/*stimme eher zu*) zu einer neuen Ausprägung (*Ablehnung* sowie *Zustimmung*) gebündelt. Damit ließ sich pro Item der Anteil an Zustimmung bzw. Ablehnung an der Gesamtzahl der Expert:innenurteile berechnen. Da die vier Ausprägungen der in der Online-Befragung verwendeten Likert-Skala nicht äquidistant sind, werden die Daten als ordinalskaliert aufgefasst (Brück & Toth, 2022; Rosebrock et al., 2019).

Die Freitextantworten der Expert:innen wurden hingegen qualitativ in Form einer inhaltlichen Kategorisierung ausgewertet. Dazu wurden alle Freitextantworten itemweise von den Autor:innen gesichtet und im Rahmen eines konsensuellen Vorgehens (Kuckartz, 2022) einer der vier folgenden Kategorien zugeordnet. Die Kategorien wurden nach einer ersten Sichtung aus den Daten heraus gebildet:

Positive Rückmeldungen (Begründung des Expert:innenurteils im Falle von Zustimmung), Kritik (Begründung des Expert:innenurteils im Falle von Ablehnung), Vorschläge zur Änderung der Itemformulierung sowie Vorschläge zur Ergänzung des Itempools.

5 Ergebnisse

5.1 Quantitative Daten aus dem Expert:innenrating

Zunächst werden die quantitativen Daten der Expert:innenurteile über die Repräsentanz und Relevanz der Items präsentiert, die in Abbildungen 3 und 4 visualisiert sind. Die Abbildungen zeigen die Anzahl zustimmender und ablehnender Expert:innenurteile pro Item in aufsteigender Reihenfolge. Die unterschiedliche Färbung der Balken ermöglicht es, die absoluten Urteilsanteile den jeweiligen Expert:innengruppen zuzuordnen (hell: Grundschullehrkräfte und Fachleitungen, dunkel: Hochschullehrer:innen). Zustimmung bzw. Ablehnung bedeutet in diesem Kontext, dass die Expert:innen (nicht) der Auffassung waren, das jeweilige Item bilde einen Aspekt kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 ab. Die in diesem Abschnitt präsentierten Ergebnisse erlauben folglich Rückschlüsse über die Inhaltsvalidität der Items.

Von den 41 Items des Lehrkräftefragebogens erhielten sechs Items die vollständige Zustimmung der Expert:innen und insgesamt 31 Items 75 % (= 18 Urteile) oder mehr Zustimmung. Von den 41 Schüler:innenitems erhielten fünf Items die vollständige Zustimmung der Expert:innen und insgesamt 30 Items 75 % (= 18 Urteile) oder mehr Zustimmung. Aus beiden Itempools erhielt kein einziges Item weniger als 50 % (= 12 Urteile) Zustimmung.

Aufgrund des Anteils an Zustimmung durch die Expert:innen können die Items in Ränge eingeteilt werden. So bilden beispielsweise alle Items mit 24 zustimmenden Urteilen Rang eins, alle Items mit 23 zustimmenden Urteilen Rang zwei usw. Beim Rangvergleich der Items zwischen Lehrkräfte- und Schüler:innenfragebogen zeigt sich, dass mit 34 Items ein Großteil der Items entweder denselben Rang belegt oder einen Rangunterschied von 1 aufweist. Bei den übrigen sieben Items fällt dieser Rangunterschied höher aus. Die fünf Items *einsprachigkeit*, *koerpersprache*, *modifikation*, *spass* und *sprachfehler* unterscheiden sich zwischen den beiden Fragebögen um zwei Ränge. Bei den beiden Items *focusonforms* und *simulation* beträgt der Rangunterschied sogar 4.

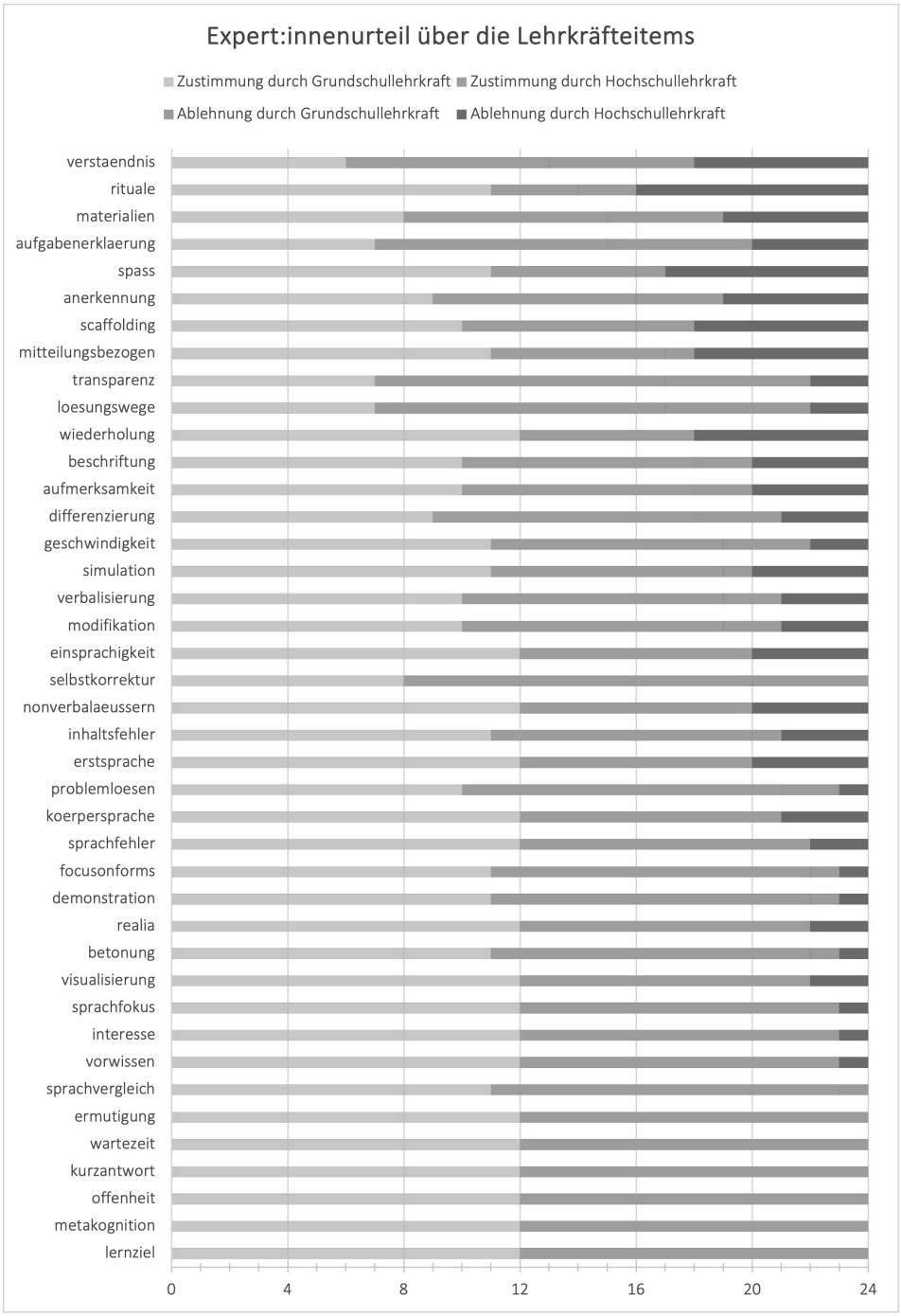


Abbildung 3: Visualisierung der Expert:innenurteile hinsichtlich der Lehrkräfteitems zu der Frage: „Bilden die Items Ihrer Meinung nach einen Aspekt kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 ab?“

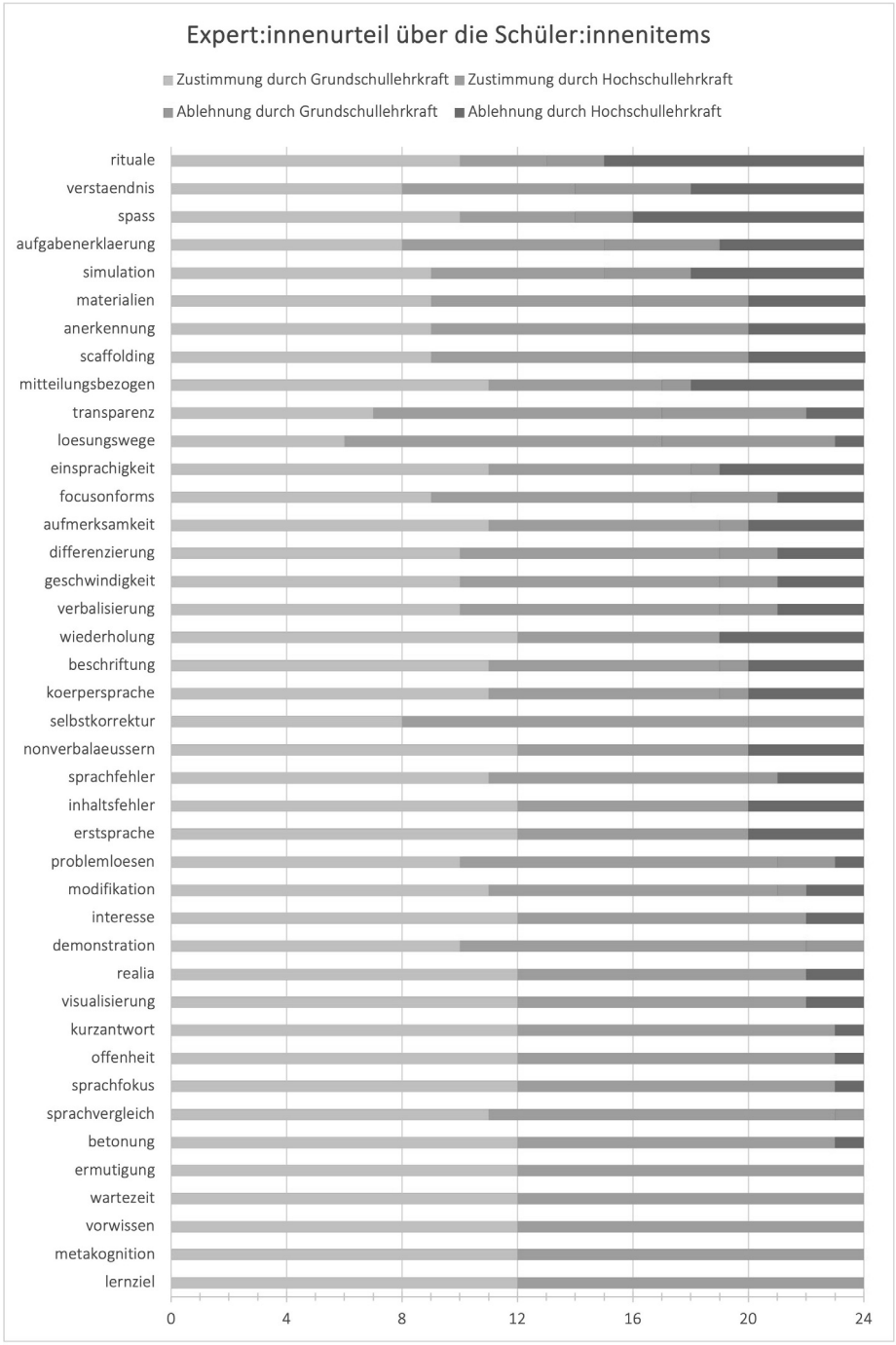


Abbildung 4: Visualisierung der Expert:innenurteile hinsichtlich der Schüler:innenitems zu der Frage: „Bilden die Items Ihrer Meinung nach einen Aspekt kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4 ab?“

5.2 Qualitative Daten aus den Expert:innenkommentaren

Nachfolgend werden die Expert:innenkommentare aus den Freitextantworten entlang der vier in Abschnitt 4.3 beschriebenen Kategorien zusammenfassend wiedergegeben. Kategorie 1 (positive Rückmeldungen) beinhaltet alle Kommentare, in denen ein zustimmendes Expert:innenurteil begründet wurde. Die Begründung des durch das Item intendierten Potenzials zu kognitiver Aktivierung erfolgte unter Bezug auf Zweitspracherwerbstheorien, pädagogisch-psychologische Lerntheorien oder die eigene Unterrichtserfahrung.

In Kategorie 2 (Kritik) wurden alle Kommentare zusammengefasst, mit denen die Expert:innen ihr ablehnendes Urteil begründeten. Diese Begründungen thematisierten folgende Aspekte:

- Gefahr konstruktirrelevanter Varianz: Vereinzelt wurde darauf hingewiesen, die Ausprägung eines Items hinge nicht ausschließlich mit dem Grad kognitiver Aktivierung zusammen. Ein Beispiel dafür ist das Item *beschriftung* ([L] „Ich beschrifte Gegenstände im Klassenraum mit englischen Wörtern oder Redewendungen.“, [S] „In meinem Klassenraum gibt es Sachen, die mit englischen Wörtern oder Sätzen beschriftet sind.“). Grundschullehrkräfte wiesen darauf hin, die Gestaltung des Klassenraums obliege in manchen Fällen der Klassenlehrkraft oder der Klassenraum sei schlicht so voll, dass eine Beschriftung einzelner Gegenstände nicht möglich sei.
- Iteminhalt differenziert nicht stark genug zwischen dem Potenzial zu kognitiver Aktivierung und kognitiver Aktivität: Bei einigen Items wurde darauf hingewiesen, es handele sich bei dem Iteminhalt lediglich um ein Oberflächenmerkmal von Unterricht, das Gelegenheitsstrukturen für das Potenzial zu kognitiver Aktivierung beschreibe. Hinsichtlich des Ausmaßes an kognitiver Aktivität sei es notwendig, den Iteminhalt in Bezug zu einer konkreten Unterrichtsaktivität zu setzen. Beispielhaft sei hier auf das Item *materialien* ([L] „Ich statte den Klassenraum mit englischsprachigen Materialien (z. B. Bilderbücher, Spiele, CDs) für die Schüler:innen aus.“, [S] „In meinem Klassenraum gibt es Bücher, Spiele oder CDs auf Englisch, die ich benutzen darf.“) verwiesen.
- Iteminhalt weist einen Mangel an unterrichtspraktischer Relevanz auf: Obwohl einzelne Teilnehmer:innen eine Passung zwischen Iteminhalt und Konstruktdefinition feststellten, beurteilten sie das Item mit Ablehnung, da der Iteminhalt selten bis gar nicht in ihrer Unterrichtspraxis vorkomme. Diese Form der Begründung wurde beispielsweise beim Item *loesungswege* ([L] „Wenn Schüler:innen unterschiedliche Formulierungen vorschlagen, um dasselbe zu sagen, diskutiere ich mit ihnen, welcher Vorschlag am angemessensten ist.“, [S] „Wenn es im Englischunterricht mehrere Antworten auf eine Frage gibt, überlegen wir als Klasse, welche Antwort die Beste ist.“) geäußert.
- Keine Passung zwischen Item und Unterrichtsqualitätsdimension kognitiver Aktivierung: Bei der Beurteilung des Items *spass* ([L] „Ich sorge dafür, dass sich die Schüler:innen wohlfühlen und Spaß beim Lernen haben.“, [S] „Ich habe Spaß am Englischunterricht und spreche gerne auf Englisch.“) wurde

beispielsweise mehrfach auf eine höhere Passung des Items mit der Unterrichtsqualitätsdimension sozio-emotionale Unterstützung hingewiesen. Vorschläge für die Zuordnung eines Items zu einer anderen Unterrichtsqualitätsdimension waren nicht Teil der Online-Befragung, sondern wurden selbstständig von den Expert:innen angeführt, basierten also auf ihrem jeweils individuellen Vorwissen.

In Kategorie 3 fallen all jene Kommentare, in denen die Expert:innen Aspekte der Itemformulierung diskutierten oder Vorschläge zur Überarbeitung der Itemformulierung verfassten. In dieser Kategorie lassen sich vier wiederkehrende Themen identifizieren. Erstens wurde dazu geraten, die Formulierungen in den Lehrkräfte- und Schüler:innenitems stärker aneinander anzupassen, um die Konsistenz der beiden jeweils analog konstruierten Items zu erhöhen. So wurde beispielsweise kritisiert, dass die Umschreibung des Wortes „Anschauungsmaterialien" aus dem Lehrkräfteitem *realia* mit dem Ausdruck „Sachen, die ich von zuhause kenne" im Item der Schüler:innen inhaltlich zu ungenau sei.

Zweitens wiesen die Expert:innen darauf hin, einige der Items seien sprachlich zu komplex. Um diesem Problem zu begegnen, wurde zunächst dazu geraten, die Schüler:innenitems so kurz und simpel wie möglich zu formulieren. Außerdem schlugen die Expert:innen vor, im Grundschulalter seltener genutzte Wörter durch bekanntere Wörter zu ersetzen (z. B. „zeigen" statt „verdeutlichen" oder „erklären"). Schließlich identifizierten die Expert:innen in wenigen Items mehrdeutige Wörter, die ambige Iteminhalte verursachten. Zur Veranschaulichung sei das Lehrkräfteitem *wiederholung* („Ich wiederhole Wörter und Redewendungen, die zum Lernziel der Unterrichtsstunde gehören.") angeführt. In diesem Item verstanden einige Expert:innen das Verb „wiederholen" im intendierten Sinne der wiederholten Aussprache. In einer Englischstunde zum Unterrichtsvorhaben *Me and my family* mit dem Lernziel Wortschatzarbeit gelte es beispielsweise als lernförderlich, wenn die Englischlehrkraft Wörter und Redewendungen aus dem Wortfeld *relatives* im Unterrichtsgespräch wiederholt verwendet. Andere Expert:innen hingegen begriffen das Verb „wiederholen" im Sinne des wiederholten Übens eines *language chunks*. Sie assoziierten damit den gehäuften Einsatz von Übungsphasen (z. B. *gap filling exercises*, Dialoge), in denen die Schüler:innen den Einsatz neu eingeführter *chunks* erproben.

Drittens kommentierten die Expert:innen den Grad der Genauigkeit der Items hinsichtlich der Sicherung des Itemverständnisses. Sie rieten einerseits dazu, weitere Items um konkrete Beispiele zu ergänzen, da die Iteminhalte sonst zu abstrakt für Schüler:innen der Jahrgangsstufe 4 seien. Gleichzeitig äußerten einige der Expert:innen Bedenken darüber, Beispiele könnten aufgrund der Vielfalt möglicher Implementationen von Iteminhalten im Englischunterricht einschränkend wirken. Beispielsweise wurde vorgeschlagen, das Schüler:innenitem *rituale* („Im Englischunterricht gibt es Sachen, die mein:e Englischlehrer:in immer gleich macht.") durch die Beschreibung einer konkreten Unterrichtssituation zu konkretisieren.

Viertens benannten die Expert:innen Probleme bezüglich der Itemreferenz. Im Falle der Lehrkräfteitems sei vereinzelt unklar, ob sich die Iteminhalte auf die Interaktion zwischen Lehrkraft und Schüler:in oder Lehrkraft und Lerngruppe bezögen. Aus den Schüler:innenitems ginge wiederum vereinzelt nicht hervor, ob die Iteminhalte mit Blick auf individuelle Schüler:innen oder die gesamte Lerngruppe zu bewerten seien. Zuletzt merkten die Expert:innen an, einige Items könnten hinsichtlich ihrer Formulierung geschärft werden, um den Unterschied zwischen lehrkräfteseitigem Unterrichts*angebot* und schüler:innenseitiger Unterrichts*nutzung* stärker herauszuarbeiten.

Kategorie 4 beinhaltet die Vorschläge der Expert:innen zur Erweiterung des Itempools um weitere Aspekte des Englischunterrichts in der Primarstufe, die sich wie folgt zusammenfassen lassen:

- Differenzierung der Items zu *corrective feedback* hinsichtlich konkreter Feedback-Strategien,
- Einbezug von für die Primarstufe typischen Unterrichtsgegenständen und Aufgabenstellungen,
- Einführung des Schriftbildes,
- Einsatz digitaler Medien,
- formatives Assessment,
- Formen der Ausspracheschulung (z. B. durch Chorsprechen, Lieder oder Sprachspiele),
- Hinzufügen weiterer Items zur Interaktion der Schüler:innen in der Zielsprache, um diesen Teilaspekt der Konstruktdefinition nicht zu vernachlässigen.

Schließlich gab es vereinzelt Expert:innen, die der Auffassung waren, der Itempool beinhalte zu viele Items zu allgemeinen Strategien des Fremdsprachenunterrichts, die nicht unmittelbar relevant für kognitive Aktivierung seien. Sie forderten die Formulierung zusätzlicher Items zum Problemlösen sowie zum Einsatz herausfordernder Aufgaben im Englischunterricht der Primarstufe. Als Beispiele führten sie offene Unterrichtssettings und Aufgabenformate wie das *task-based language teaching* an sowie Gelegenheiten zum selbstständigen Lernen.

6 Ergebnisdiskussion

Nachfolgend werden die präsentierten Ergebnisse der Online-Expert:innenbefragung diskutiert und auf dieser Grundlage Implikationen für die Weiterentwicklung des Fragebogens zur Erfassung kognitiver Aktivierung im Englischunterricht der Primarstufe herausgearbeitet. Eine eindeutige Interpretation der Expert:innenurteile zu den Schüler:innen- und Lehrkräfteitems ist insofern herausfordernd, als dass für die Wertung des Vorliegens von Inhaltsvalidität – anders als beispielsweise für Reliabilitätskoeffizienten – keine Normwerte existieren. In Studien, die ebenfalls einen quantitativen Zugang zur Annäherung an Inhaltsvalidität wählen (vgl. Abschn. 2.3), formulieren die Autor:innen eigene Kriterien für die Item-

selektion, die vor dem Hintergrund des jeweiligen Forschungsprojekts plausibel erscheinen. Diesem Vorgehen folgend wurden für die vorliegende Teilstudie vor Durchführung der Online-Befragung folgende Kriterien formuliert:

- Kriterium K1: Items, die von mindestens der Hälfte der Expert:innen (= 12 Urteile) ein ablehnendes Urteil erhalten, gelten nicht als ausreichend inhaltsvalide und werden eliminiert.
- Kriterium K2: Items, die von maximal 15 % der Expert:innen (= 4 Urteile) ein ablehnendes Urteil erhalten, gelten als ausreichend inhaltsvalide und werden ohne inhaltliche Änderungen beibehalten.
- Kriterium K3: Items, die von mindestens 15 % bis maximal 50 % der Expert:innen ein ablehnendes Urteil erhalten, werden nach einer erneuten Prüfung entweder eliminiert oder auf Grundlage der Expert:innenkommentare überarbeitet.

Die Revisionsgrenze von 15 % wurde im Vergleich zu Studien mit einem ähnlichen methodischen Vorgehen aufgrund der in Abschnitt 2.1 geschilderten Unschärfe des Konstrukts kognitiver Aktivierung bewusst strenger gewählt. Anhand dieser Kriterien lässt sich vorsichtig schlussfolgern, dass die theoriebasierte Itemkonstruktion in einen inhaltsvaliden Itempool für beide Fragebögen resultierte. Keines der Schüler:innen- oder Lehrkräfteitems wurde durch K1 eliminiert. 23 Lehrkräfteitems und 21 Schüler:innenitems wurden gemäß K2 inhaltlich unverändert beibehalten. Dieser Anteil entspricht in beiden Fällen mehr als der Hälfte des Itempools. Damit unterlagen im Anschluss an die Online-Expert:innenbefragung 18 Lehrkräfteitems und 20 Schüler:innenitems einer erneuten Prüfung und Überarbeitung.

Die aus den Expert:innenurteilen hervorgegangenen Rangunterschiede zwischen den jeweils analogen Schüler:innen- und Lehrkräfteitems könnten ein Indiz für perspektivenspezifische Aspekte kognitiver Aktivierung liefern (Fauth et al., 2020; Lindl et al., in diesem Band). Diese Rangunterschiede können so interpretiert werden, dass die Expert:innen das jeweilige Item hinsichtlich kognitiver Aktivierung für Lehrkräfte relevanter halten als für Schüler:innen oder umgekehrt. So merkten einige Expert:innen zu Lehrkräfteitems mit einem starken Fokus auf das Lehrkräfteverhalten oder didaktische Aspekte an, der Iteminhalt sei konstruktirrelevant für den Schüler:innenfragebogen, da Schüler:innen nicht über genügend Informationen und Expertise verfügten, um das Item zu beantworten. Auskunft über die Gültigkeit dieser Vermutung werden Ergebnisse weiterer Validierungs- und Pilotierungsstudien liefern.

Zudem wurde bei der Itemkonstruktion darauf geachtet, Unterrichtsmerkmale zu fokussieren, die *Potenziale* zu kognitiver Aktivierung repräsentieren, und diese Unterrichtsmerkmale – sofern möglich – anhand konkreter Szenarien zu beschreiben. Dies bedeutet, dass die Lehrkräfteitems zu kognitiver Aktivierung mehrheitlich selbstreferent und -evaluativ sind, wohingegen die Schüler:innenitems mehrheitlich fremdreferent und -evaluativ sind. Vergleichsweise hohe Rangunterschiede treten primär bei solchen Items auf, die davon eine Ausnahme

bilden. Ein Beispiel dafür ist das Schüler:innenitem *focusonforms* („Wenn ich etwas auf Englisch falsch sage, erklärt mir mein:e Englischlehrer:in den Fehler. Ich denke dann ganz genau über den Fehler nach."), das aufgrund der Expert:innenurteile vier Ränge über dem zugehörigen Lehrkräfteitem („Wenn Schüler:innen im Gespräch einen sprachlichen Fehler machen, nutze ich das als Anlass, ihnen diesen Fehler zeitnah bewusst zu machen und zu erklären.") liegt. Dieses Item fordert sowohl von Lehrkräften als auch von den Schüler:innen, ein kombiniertes Urteil über beide Personengruppen abzugeben.

Ferner könnten die Rangunterschiede in der Expertise der befragten Expert:innen begründet sein, die auf unterschiedlichen Wissensformen (Vogel, 2019) basiert. Die Hochschullehrkräfte verfügen allein aufgrund ihrer Disziplinzugehörigkeit über variierendes Fachwissen zu kognitiver Aktivierung, wohingegen sich die Grundschullehrkräfte in ihrem unterrichtspraktischem Erfahrungswissen unterscheiden. Zudem ist davon auszugehen, dass sich die Hochschullehrkräfte vorwiegend an wissenschaftlichem Wissen orientieren, das seitens der Grundschullehrkräfte während der Berufspraxis durch pädagogisches Alltags- und Professionswissen angereichert wurde. Betrachtet man die Expert:innenurteile separat nach Personengruppen, wird erkennbar, dass die Grundschullehrkräfte deutlich mehr Items (z. B. *wiederholung*, *nonverbalaeussern* oder *realia*) ausschließlich zustimmend beurteilten als die Hochschullehrkräfte. Umgekehrt ist dies nur bei drei Items der Fall (*sprachvergleich*, *demonstration* und *selbstkorrektur*). Eine Verzerrung aufgrund Zustimmung durch soziale Erwünschtheit ist seitens der Grundschullehrkräfte eher unwahrscheinlich, da sie im Rahmen der Online-Befragung nicht dazu aufgefordert waren, ihren eigenen Unterricht zu beurteilen. Auch wenn es nicht der Regelfall ist, scheint bei einigen Items durchaus Uneinigkeit zwischen beiden Personengruppen zu herrschen. Auch in den Expert:innenkommentaren aus den Freitextantworten dokumentieren sich personengruppenspezifische Merkmale. Die Hochschullehrkräfte – insbesondere solche, die sich der empirischen Bildungsforschung zugehörig fühlen – wählen mehrheitlich einen komparativen Ansatz und diskutieren die Items im Kontext der Unterrichtsqualitätsforschung. Sie verweisen beispielsweise auf andere Unterrichtsqualitätsdimensionen des Syntheseframeworks (Praetorius et al., 2020) oder vergleichen die neu konstruierten Items mit bereits veröffentlichten Items zur Erfassung kognitiver Aktivierung. Die Grundschullehrkräfte schilderten zur Begründung ihrer Itemurteile häufig Episoden des Englischunterrichts und leiteten daraus die Relevanz des jeweiligen Items ab. Obwohl diese unterrichtspraktischen Erfahrungen höchst subjektiv sind und nicht unmittelbar verallgemeinert werden sollten, liefern sie doch wichtige Informationen über die ökologische Validität der Items.

Die Vorschläge der Expert:innen zur Ergänzung des Itempools aus den Freitextantworten wurden einzeln geprüft. Die Prüfung erfolgte vor dem Hintergrund der Konstruktspezifikation kognitiver Aktivierung, der zur Konstruktion des Itempools gesichteten Literatur sowie dem Zweck der zu entwickelnden Fragebögen. Der Vorschlag, die Items zu *corrective feedback* zu differenzieren, wurde

beispielsweise bei der Revision des Itempools berücksichtigt. Im Entwurf des Itempools waren drei vergleichsweise allgemein formulierte Items zum Umgang der Lehrkraft mit Schüler:innenfehlern enthalten, die sich auf die Korrektur von Sach- und Sprachfehlern sowie die Anregung der Schüler:innen zur Selbstkorrektur sprachlicher Fehler bezogen. Die Expert:innen äußerten diesbezüglich mehrheitlich, die existierenden Items seien zu undifferenziert, da nicht alle Formen von *corrective feedback* gleichermaßen kognitiv aktivierend wirken. Diese Annahme unterschiedlicher Effekte verschiedener Feedback-Strategien wird durch die empirische Befundlage der Zweitspracherwerbsforschung gestützt (z. B. *immediate vs. delayed feedback*, *implicit vs. explicit feedback*; Li, 2010; Lyster et al., 2013). Im Zuge der Revision wurden sechs neue Items (*explicit correction, recast, request, metalinguistic cue, repetition, elicitation*; Oliver & Adams, 2021) formuliert, die empirisch identifizierte Feedback-Strategien des Fremdsprachenunterrichts abbilden und so dazu dienen, den Umgang der Lehrkraft mit Schüler:innenfehlern differenzierter zu erfassen.

Andere Ergänzungsvorschläge, wie beispielsweise der Einbezug von für die Primarstufe typischen Unterrichtsgegenständen und Aufgabenstellungen, wurden nach einer kritischen Würdigung und Prüfung verworfen. In diesem Fall war die Entscheidung darin begründet, dass die zu entwickelnden Skalen bzw. Inventare im weiteren Verlauf des Promotionsprojekts zur Untersuchung von Zusammenhängen mit weiteren Konstrukten möglichst breit einsetzbar sein sollen. Dieses Ziel wäre durch eine Ausrichtung der Items auf konkrete Unterrichtsgegenstände, die möglicherweise nur zu einem bestimmten Zeitpunkt im Schuljahr unterrichtet werden, eingeschränkt worden. Dennoch ist im Forschungsdiskurs zu Unterrichtsqualität eine Auseinandersetzung mit der Beziehung zwischen den Inhalten und der Qualität von Unterricht unter dem Begriff der Lerngegenstandsorientierung – auch bezüglich der Erfassung von Unterrichtsqualität – erkennbar: „In diesem Zusammenhang erscheint es zudem lohnend zu diskutieren, welche Dimensionen von Unterrichtsqualität für *alle* zentralen Ziele, Inhalte und Methoden hohe Relevanz haben und welche lediglich oder vornehmlich für *bestimmte* Ziel-Inhalts-Methoden-Konfigurationen bedeutsam sind“ (Praetorius et al., 2020, S. 429; Hervorhebung im Original). Es erscheint folglich plausibel, die pilotierten Fragebogenitems in zukünftigen Forschungsprojekten entlang von Lerngegenständen des Englischunterrichts in der Primarstufe zu konkretisieren und auf ihre Gültigkeit zu prüfen.

Abschließend werden Einschränkungen der vorliegenden Studie und Handlungsalternativen dargelegt. Für einen quantitativen Zugang, der zur Beurteilung der Inhaltsvalidität gewählt wurde, ist die Anzahl der Expert:innen relativ gering. Zudem stellen die Teilnehmer:innen aufgrund der Freiwilligkeit der Teilnahme an der Online-Befragung eine Positivauswahl dar. Eine umfangreichere Stichprobe hätte die Zuverlässigkeit des Itemratings sicherlich gesteigert; gleichzeitig ist die Anzahl der Expert:innen vergleichbar mit anderen Studien, in denen ein ähnliches methodisches Vorgehen gewählt wurde (Jenßen et al., 2015; Wibowo & Heemsoth, 2019). Des Weiteren wäre es denkbar gewesen, die Studie in einem

anderen Format wie einer Delphi-Befragung oder qualitativen Expert:inneninterviews durchzuführen. Während diese beiden Formate vermutlich eine intensivere inhaltliche Auseinandersetzung der Expert:innen mit dem Itempool ermöglicht hätten, führte die Entscheidung für eine Online-Befragung zu einer Abbildung mehrerer Perspektiven. Expert:inneninterviews in ähnlichem Umfang (z. B. Prusse-Hess & Prusse, 2018) hätten zwar auch unterschiedliche Perspektiven berücksichtigen können, wären jedoch sowohl für den Forschenden als auch die teilnehmenden Expert:innen deutlich ressourcenintensiver. Darüber hinaus bestand mit den Freitextantworten in der Online-Befragung die Gelegenheit zur Angabe von Begründungen, die von den Expert:innen in erfreulichem Ausmaß wahrgenommen wurde. Zuletzt ist anzumerken, dass die eigens formulierte Konstruktspezifikation kognitiver Aktivierung den Expert:innen ausschließlich schriftlich mitgeteilt wurde. Unklarheiten oder Missverständnisse seitens der Expert:innen, die möglicherweise zu invaliden Itembeurteilungen führen können, sind deshalb nicht auszuschließen.

7 Fazit

Auf Grundlage der Ergebnisse der Online-Expert:innenbefragung erwies sich der theoriebasiert konstruierte Itempool als solide Ausgangsbasis für die Konstruktion des Schüler:innen- und Lehrkräftefragebogens zu Erfassung kognitiver Aktivierung im Englischunterricht der Jahrgangsstufe 4. Dennoch enthielt die Expert:innenkritik konstruktive Vorschläge zur Ergänzung und Überarbeitung einzelner Items. Sie diente somit als Korrektiv für die Autor:innen, die den initialen Itempool eigens formulierten. Dieser wird im Rahmen der nächsten Teilstudie einer Überprüfung des Itemverständnisses mittels kognitiver Interviews mit Grundschullehrkräften und -schüler:innen unterzogen, um den Validierungsprozess fortzuführen (Guttke & Porsch, in Begutachtung).

Es lässt sich schlussfolgern, dass die Inhaltsvalidierung durch eine Online-Expert:innenbefragung einen produktiven, effizienten und qualitätssichernden Arbeitsschritt in der Instrumentenentwicklung bildet. Aus diesen Gründen ist es wünschenswert, dass die Sicherung der Inhaltsvalidität neu entwickelter Instrumente in zukünftigen Arbeiten innerhalb der Fremdsprachenforschung und darüber hinaus stärkere Berücksichtigung findet. Schließlich leistet die Studie einen Beitrag zu der gemeinsamen Aufgabe von Bildungswissenschaft und Fachdidaktik, Befunde zur Fachlichkeit von Unterrichtsqualität zu integrieren (Lindl et al., in diesem Band). So ist es für das Fach Englisch primär die Zweitspracherwerbsforschung, weniger die fachdidaktische Forschung, die Belege für potenziell kognitiv aktivierende Merkmale des Fremdsprachenunterrichts hervorgebracht hat. Es ist an der Zeit, die vielfältigen konzeptionellen Überlegungen der Fremdsprachendidaktik, welche die Unterrichtsgestaltung zweifelsohne bereichern, ebenfalls empirisch auf ihre Qualität zu prüfen. Zudem wird deutlich, dass die seitens der bildungswissenschaftlichen Forschung eingesetzten Instrumente zur Messung der

Unterrichtsqualitätsdimension kognitive Aktivierung lediglich eingeschränkt für fremdsprachendidaktische Untersuchungen übertragen lassen. Notwendig sind fachspezifische Anpassungen bzw. Entwicklungen von Items, die den Gegenstand angemessen zu erfassen vermögen. Die für die Fremdsprachenvermittlung bestehende Besonderheit, dass im Unterricht die Sprache Lerngegenstand und Medium zugleich darstellt, sollte dazu angemessen berücksichtigt werden. Die vorliegende Studie beschreibt einen Schritt zur Entwicklung solcher Items und lässt die Empfehlung zu, dass eine Expert:innenbefragung für die Entwicklung neuer Items notwendig ist, um inhaltsvalide Messinstrumente zu erstellen.

Literatur

Begrich, L., Praetorius, A.-K., Decristan, J., Fauth, B., Göllner, R., Herrmann, C., Kleinknecht M., Taut, S. & Kunter, M. (2023). Was tun? Perspektiven für eine Unterrichtsqualitätsforschung der Zukunft. *Unterrichtswissenschaft, 51*, 63–97. https://doi.org/10.1007/s42010-023-00163-4

Brück, N. & Toth, C. (2022). *Studienbuch Operationalisierungen*. Berlin: Springer. https://doi.org/10.1007/978-3-658-30239-9

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Berlin: Springer. https://doi.org/10.1007/978-3-642-41089-5

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive. Zusammenhänge und Vorhersage von Lernerfolg. Z*eitschrift für Pädagogische Psychologie, 28*(3), 127–137. https://doi.org/10.1024/1010-0652/a000129

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K. & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik, 66* (Beiheft), 138–155. https://doi.org/10.3262/ZPB2001138

Gass, S. M., Behney, J. & Plonsky, L. (2020). *Second language acquisition: An introductory course.* Routledge. https://doi.org/10.4324/9781315181752

Göllner, R., Wagner, W., Klieme, E., Lüdtke, O., Nagengast, B. & Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven. In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments* (S. 63–82). Berlin: Bundesministerium für Bildung und Forschung.

Gruehn, S. (2000). *Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung.* Münster: Waxmann.

Grünkorn, J., Klieme, E. & Stanat, P. (2018). Bildungsmonitoring und Qualitätssicherung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, C. K. Spieß, H. Solga & K. Zimmer (Hrsg.), *Das Bildungswesen in Deutschland: Bestand und Potentiale* (S. 263–298). Bad Heilbrunn: UTB/Klinkhardt.

Guttke, J. (2023). Kognitive Aktivierung im Fremdsprachenunterricht: Ein systematisches Review von Forschungsarbeiten aus dem deutschsprachigen Raum. *Zeitschrift für Fremdsprachenforschung, 34*(2),145–175.

Guttke, J. & Porsch, R. (in Begutachtung). *Kognitive Interviews als Methode der Inhaltsvalidierung von Fragebogenitems zur Erfassung kognitiver Aktivierung im Englischunterricht der Grundschule*.

Helmke, A. (2014). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (5. Aufl.). Seelze-Velber: Klett/Kallmeyer.

Jenßen, L., Dunekacke, S. & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *Kompetenzen von Studierenden* (S. 11–31). Weinheim/Basel: Beltz Juventa.

Kersten, K., Bruhn, A.-C., Ponto, K., Böhnke, J. & Greve, W. (2018). Teacher input observation scheme (TIOS). *Studies on Multilingualism in Language Education*, 4. Universität Hildesheim.

Kirchhoff, P. (2016): Was sollte eine gute Englischlehrkraft wissen? Über die Auswahl von Items im FALKO-E Test zum fachspezifischen Professionswissen. In M. Legutke & M. Schart (Hrsg.), *Fremdsprachendidaktische Professionsforschung: Brennpunkt Lehrerbildung* (S. 75–98). Tübingen: Narr.

Kirchhoff, P. (2017). FALKO-E: Fachspezifisches professionelles Wissen von Englischlehrkräften. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff & R. Mulder (Hrsg.), *FALKO Fachspezifische Lehrerkompetenzen* (S. 113–152). Münster: Waxmann.

Klauer, K. J. (1984). Kontentvalidität. *Diagnostica, 30*(1), 1–23.

Kleickmann, T., Steffensky, M. & Praetorius, A.-K. (2020). Quality of Teaching in Science Education. In A.-K. Praetorius, J. Grünkorn & E. Klieme (Hrsg.), *Empirische Forschung zu Unterrichtsqualität: Theoretische Grundfragen und quantitative Modellierungen* (S. 37–55). Weinheim: Beltz Juventa.

Klieme, E. (2022). Unterrichtsqualität. In M. Harring, C. Rohlfs & M. Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (S. 411–426). Münster: Waxmann.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS – Impulse für Schule und Unterricht* (S. 43–58). Bonn: Bundesministerium für Bildung und Forschung.

Kuckartz, U. & Rädiker, S. (2022). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung.* Weinheim/Basel: Beltz/Juventa.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–235. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Lindl, A., Ehrich, P., Gutsmiedl, M., Rader, M., Simböck, L., Gürtner, M., Böhringer, S., Krämer, A., Kirchhoff, J. & Frei, M. (2024). Und wo bleibt die Ästhetik? – Betrachtungen zu einer weiteren Dimension von Unterrichtsqualität aus interdisziplinärer Perspektive. In M. Hemmer, C. Angele, C. Bertsch, S. Kapelari, G. Leitner & M. Rothgangel (Hrsg.), *Fachdidaktik im Zentrum von Forschungstransfer und Transferforschung* (S. 371–388). Münster: Waxmann.

Lipowsky, F. & Bleck, V. (2019). Was wissen wir über guten Unterricht? – Ein Update. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität: Konzepte und Bilanzen gelingenden Lehrens und Lernens* (S. 219–249). Münster: Waxmann.

Lipowsky, F. & Hess, M. (2019). Warum es manchmal hilfreich sein kann, das Lernen schwerer zu machen. In K. Schöppe & F. Schulz (Hrsg.), *Kreativität & Bildung – Nachhaltiges Lernen* (S. 77–132). München: kopaed.

Lipowsky, F. (2020). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 69–118). Springer. https://doi.org/10.1007/978-3-662-61403-7_4

Loewen, S. (2020). *Introduction to instructed second language acquisition*. Routledge. https://doi.org/10.4324/9781315616797

Lyster, R., Saito., K. & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

Maier, U., Kleinknecht, M., Metz, K. & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerbildung, 28*(1), 84–96. https://doi.org/10.36950/bzl.28.1.2010.9798

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Moosbrugger, H. & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer. https://doi.org/10.1007/978-3-662-61532-4

Oliver, R. & Adams, R. (2021). Oral corrective feedback. In H. Nassaji & E. Kartchava (Hrsg.), *The Cambridge handbook of corrective feedback in second language learning and teaching* (S. 187–206). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108589789.010

Porsch, R. & Wilden, E. (2022). Teaching English out-of-field in primary school: Differences in professional characteristics and effects on instructional quality. In L. Hobbs & R. Porsch (Hrsg.), *Out-of-field teaching across teaching disciplines and contexts* (S. 117–134). Singapur: Springer. https://doi.org/10.1007/978-981-16-9328-1_6

Praetorius, A.-K. & Gräsel, C. (2021). Noch immer auf der Suche nach dem heiligen Gral: Wie generisch oder fachspezifisch sind Dimensionen der Unterrichtsqualität? *Unterrichtswissenschaft*, *49*, 167–188. https://doi.org/10.1007/s42010-021-00119-6

Praetorius, A.-K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B. & Nehring, A. (2020). Unterrichtsqualität in den Fachdidaktiken im deutschsprachigen Raum – zwischen Generik und Fachspezifik. *Unterrichtswissenschaft, 48*, 409–446. https://doi.org/10.1007/s42010-020-00082-8

Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *Mathematics Education*, *50*, 407–426. https://doi.org/10.1007/s11858-018-0918-4

Prusse-Hess, B. & Prusse, M. (2018). *Wirksamer Englischunterricht*. Schneider Verlag Hohengehren.

Rieser, S. & Decristan, J. (2023). Kognitive Aktivierung in Befragungen von Schülerinnen und Schülern: Unterscheidung zwischen dem Potential zur kognitiven Aktivierung und der individuellen kognitiven Aktivierung. *Zeitschrift für Pädagogische Psychologie*, 1–15. https://doi.org/10.1024/1010-0652/a000359

Rosebrock, A., Schlosser, S., Höhne J. & Kühnel, S. M. (2019). Einflüsse unterschiedlicher Formen der Verbalisierung von Antwortskalen auf das Antwortverhalten von Befragungspersonen. In N. Menold & T. Wolbring (Hrsg.), *Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente* (S. 65–102). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-24517-7_3

Schreyer, P., Wemmer-Rogh, W., Herbert, B. & Praetorius, A.-K. (2022). *Kognitive Aktivierung Kognitive Aktivierung: Ein systematischer Überblick.* Vortrag auf der 9. GEBF-Tagung, Bamberg.

Senkbeil, M., Ihme, J. & Wittwer, J. (2013). Entwicklung und erste Validierung eines Tests zur Erfassung technologischer und informationsbezogener Literacy (TILT) für Jugendliche am Ende der Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft*, *16*, 671–791. https://doi.org/10.1007/s11618-013-0446-5

Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK] (2023). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Ersten Schulabschluss und den Mittleren Schulabschluss.* Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2023/2023_06_22-Bista-ESA-MSA-ErsteFremdsprache.pdf

Vieluf, S., Praetorius, A.-K., Rakoczy, K., Kleinknecht, M. & Pietsch, M. (2020). Angebots-Nutzungs-Modelle der Wirkweise des Unterrichts: Ein kritischer Vergleich verschiedener Modellvarianten. In A.-K. Praetorius, J. Grünkorn & E. Klieme (Hrsg.), *Empirische Forschung zu Unterrichtsqualität: Theoretische Grundfragen und quantitative Modellierungen* (S. 63–80). Weinheim: Beltz Juventa. https://doi.org/10.3262/ZPB2001063

Weber, K.-A., Friege, G. & Scholz, R. (2020). Quantenphysik in der Schule – Was benötigen Lehrkräfte? Ergebnisse einer Delphi-Studie. *Zeitschrift für Didaktik der Naturwissenschaften*, *26*, 173–190. https://doi.org/10.1007/s40573-020-00119-6

Wibowo, J. & Heemsoth, T. (2019). Fachdidaktisches Wissen von Sportlehrer*innen testen: Überlegungen zur Inhaltsvalidität. *Zeitschrift für sportpädagogische Forschung*, *2*, 88–108. https://doi.org/10.5771/2196-5218-2019-2-88

Wilden, E., Porsch, R., Guttke, J. & Wellmans, L. (eingereicht). Das Potenzial zur kommunikativ-kognitiven Aktivierung im Englischunterricht – Befunde einer Analyse von Lernaufgaben aus Lehrwerken der Grundschule.