

Messinstrument zur Analyse und Kategorisierung von MINT-Aufgaben – Konstruktion und Validierung

Sebastian Stuppan^{a,b}, Markus Wilhelm^{a,b}, Katrin Bölsterli Bardy^a, Roland Künzle^a

^aPädagogische Hochschule Luzern

^bPädagogische Hochschule Heidelberg, Deutschland

sebastian.stuppan@phlu.ch

Abstract

In einem kompetenzorientierten Unterricht haben Aufgaben eine entscheidende Bedeutung. Mit Hilfe von Aufgabensets werden Gelegenheiten geschaffen, damit Lernende ihre Kompetenzen individuell entwickeln und darin gefördert werden. Im Forschungsvorhaben wird ein Messinstrument zur Analyse und Kategorisierung von MINT Aufgaben erarbeitet. Dieser Artikel beschreibt und diskutiert die Konstruktion und Validierung des Aufgaben-Analyse-Instruments (AAI).

Keywords

Validierungsstrategie; Messinstrument; MINT - Aufgaben; Kategorisierung; Skalenentwicklung.

Einleitung

Aufgaben können im schulischen Lernen als Dreh- und Angelpunkt betrachtet werden. Entsprechend steuern sie den Lehr-Lernprozess (Luthiger et al., 2018, S. 11; Reusser, 2014). Damit der in den Lehrplänen verlangte Kompetenzaufbau der Lernenden erreicht werden kann, treten u.a. Sets von aufeinander aufbauenden Aufgaben in den Fokus (Büchter & Leuders, 2006). Ein Beispiel, wie ein solches Aufgabenset aufgebaut sein kann, ist das Luzerner Modell zur Entwicklung kompetenzfördernder Aufgabensets: LUKAS-Modell (Luthiger et al., 2018). In einzelnen Fachdidaktiken und in der Allgemeinen Didaktik wurden verschiedene Kategoriensysteme zur Analyse oder Kategorisierung einzelner Aufgaben entwickelt und teilweise erprobt (Blömeke et al., 2006; Bölsterli Bardy & Wilhelm, 2018; Jordan et al., 2006; Luthiger et al., 2018; Maier et al., 2010; Neubrand, 2002). Da momentan empirisch getestete Instrumente zur Analyse von MINT-Aufgabensets nach dem LUKAS-Modell fehlen (Luthiger et al., 2018, S. 58), erzielt dieses Dissertationsvorhaben, die Entwicklung und Validierung eines Messinstruments auf der Grundlage bestehender Kategoriensysteme. Die Kategorien des Aufgaben-Analyse-Instruments wurden aus bestehenden Aufgabenklassifikationen (Luthiger et al., 2018; Wilhelm et al., 2014) abgeleitet und zu Skalen mit entsprechenden Subskalen weiterentwickelt. Zur Validierung des Messinstruments wurden Aufgaben aus dem Primarschulprojekt «MINT unterwegs» genutzt, welche auf dem LUKAS-Modell basieren.

Theoretischer Hintergrund

In den meisten pädagogisch-didaktischen Fachbereichen haben Aufgaben eine zentrale Bedeutung, da sie ein wesentliches Steuerungsinstrument in Lehr-Lern-Situationen und wichtige Elemente des Lernweges darstellen. Deshalb schreiben ihnen Müller und Helmke (2008, S. 32) für die Unterrichtsentwicklung und Unterrichtsqualität eine zentrale Rolle zu. Gemäss Reusser (2014) sind Aufgaben speziell in heterogenen Lerngruppen relevant. Während Lernaufgaben dem Aufbau von fachlichen und überfachlichen Kompetenzen dienen, sind Leistungsaufgaben zur Überprüfung von Bildungsstandards oder der Evaluation der Zielerreichung dienlich (Reusser, 2014, S. 79). Folglich ist die Analyse von Aufgaben ein bedeutendes Feld für das Verständnis und die Förderung schulischer Lehr-Lernprozesse. In den klassischen Didaktikmodellen der 60er- und 70er-Jahre des vergangenen Jahrhunderts spielen Aufgaben noch kaum eine Rolle (Maier et al., 2014, S. 35f.). Inzwischen sind Aufgabenstellungen in der empirischen Unter-

richtsforschung sowie in der Forschung der meisten Fachdidaktiken ein zentraler Gegenstand geworden. So setzen sich u.a. Adamina (2010), Bölsterli Bardy und Wilhelm (2018), Büchter und Leuders (2016), Jordan et al. (2006), Neubrand (2002) sowie Winkler (2011) mit Systemen zur Klassifikation von Aufgaben aus fachdidaktischer Perspektive auseinander. In der Allgemeinen Didaktik beschäftigen sich z.B. Blömeke et al. (2006), Maier et al. (2010) sowie Luthiger et al. (2018) mit Kategoriensystemen und der Aufgabenanalyse.

Aufgrund der hohen Relevanz von Aufgaben, sind auch bei Lehramtsstudierenden die didaktische Funktion und Passung von Aufgaben in den Vor- und Nachbereitungen ihrer Praktika zentral. In der Weiterbildung von Lehrpersonen ist die Frage nach der «guten» Aufgabe ein Thema. Zudem basieren digitalisierte und individualisierte Lernangebote im Wesentlichen auf einer Zusammenstellung von qualitativ hochwertigen Aufgaben (Maier et al., 2014, S. 36). All diesen skizzierten Anwendungsbereichen ist gemein, dass sie eine differenzierte Analyse und Bewertung von Aufgabenstellungen voraussetzen. Als mögliche Hilfe kann das LUKAS-Modell der Arbeitsgruppe um Luthiger (2018) dienen. Dem LUKAS-Modell zugrunde liegt die Idee, dass Aufgaben innerhalb des Lernprozesses einer Reihenfolge unterliegen und als Gesamtheit den gewünschten Kompetenzaufbau ermöglichen. Jede einzelne Aufgabe hat somit im Unterrichtsetting eine didaktische Funktion (z.B. Aufbau von Kompetenzen, Vertiefung, Ausdifferenzierung sowie Variantenbildung von Kompetenzen, etc.) und ist geprägt von Merkmalen (Skalen), die wiederum in unterschiedlichen Ausprägungen (Subskalen) vorliegen.

Fragestellung

Mit dem bestehenden LUKAS-Kategoriensystem können Aufgaben nur textbasiert einzelnen Ausprägungen zugeordnet werden, wie beispielsweise die Befunde von Reinfried (2016) zeigen. In der Studie werden aus bestehenden Kategoriensystemen Items für das Aufgaben-Analyse-Instrument entwickelt und folgende Fragestellungen untersucht: 1. Stimmen die Items mit den theoretischen Konstrukten überein? 2. Wie zuverlässig können Rater:innen, MINT-Aufgaben mit den Items beurteilen? 3. Lassen sich mit den Items Skalen und entsprechende Subskalen für das Aufgaben-Analyse-Instrument bilden? Es wird erwartet, dass sich die theoretischen Konstrukte mit einer *Closed Card Sorting Methode* korrekt zuordnen lassen. Mittels explorativer Faktorenanalysen sollte eine sparsame und eindeutige Faktorenlösung gefunden werden. Dies aus dem Grund, weil sich die theoretischen Konstrukte inhaltlich deutlich voneinander unterscheiden.

Methode

Für die Konstruktion und Validierung des Aufgaben-Analyse-Instruments wurden acht Schritte durchlaufen. Dabei wurden die von Bühner (2006), Busker (2014), Döring und Bortz (2016) sowie Mummendey und Grau (2014) beschriebenen Konstruktions- und Validierungsstrategien berücksichtigt. Die einzelnen Methoden der Konstruktion des theoriebasierten und empirisch geklärten Itempools des Messinstruments werden in den Teilschritten 1 bis 4 und die Validierungen des Messinstruments in den Teilschritten 5 bis 8 im kommenden Abschnitt erläutert.

1. **Konzeptspezifikation:** In einem ersten Teilschritt wurden theoriebasiert die Konstrukte und Qualitätskriterien von Aufgaben studiert und eingegrenzt.
2. **Operationalisierung:** In einem zweiten Teilschritt erfolgte mithilfe der theoretischen Grundlage eine Definition der Konstrukte und deren Unter Aspekte. Aus den erarbeiteten Definitionen wurden Itemtexte formuliert und eine 5-stufige Ratingskala von «stimmt nicht» bis «stimmt sehr» ausgearbeitet.
3. **Expert:innenbefragungen:** Mithilfe von Expert:innen-Befragungen in Anlehnung an die Delphi-Methode (Döring & Bortz, 2016, S. 420f.) wurde in einem nächsten Teilschritt der gesamte Itempool verfeinert und überarbeitet. Die Autor:innen des LUKAS-Modells fungierten als Expert:innenteam ($N = 4$, davon 50% weiblich). Alle Expert:innen verfügen über einen fachwissenschaftlichen und didaktischen Hintergrund im Bereich Natur, Mensch, Gesellschaft (NMG).

4. Expert:innen-Pretest und Revision: Zusammen mit dem Team der LUKAS-Expert:innen wurde ein Testmanual erstellt und anhand von drei Aufgabenratings pilotiert. Das Testmanual erhöht die Objektivität der Ratings (Döring & Bortz, 2016, S. 442). Die Aufgabenbeurteilungen der Rater erfolgten in der ersten Runde unabhängig voneinander. Bei der anschließenden Besprechung wurden Unstimmigkeiten diskutiert, bis eine konsensuelle Beurteilung erreicht war.
5. Itempool den theoretischen Konstrukten zuordnen: Die Inhaltsvalidität erfolgte, wie Döring und Bortz (2016, S. 446f.) vorschlugen, theoretisch-argumentativ und gestützt durch weitere Urteile von Fachpersonen. Die Fachpersonen ($N = 12$, davon 50% weiblich, alle Personen verfügen mindestens über einen Masterabschluss, vier davon über ein Doktorat) aus den MINT Fächern kannten den theoretischen Hintergrund der ausgearbeiteten Konstrukte. Sie hatten mittels einer digitalen *Closed Card Sorting Methode*, jedes Item aus dem Itempool nach inhaltlichen Überlegungen einem Unteraspekt eines Konstrukts zugeordnet (Schilb, 2005; Wandke et al., 2009).
6. Rater:innenschulung Zum Analysieren der Aufgaben durchliefen zwei Fachdidaktiker:innen eine mehrtägige Rater:innenschulung. Dabei wurden u.a. dasselbe Testmanual sowie dieselben drei Aufgaben aus dem Expert:innen-Rating verwendet. Durch die Rater:innenschulung wurde eine hohe Interrater-Reliabilität und Objektivität der Aufgabenbeurteilung mit dem Messinstrument angestrebt (Kuckartz, 2018, S. 22).
7. Aufgaben Beurteilung: Die geschulten Fachdidaktiker:innen ($N = 2$) beurteilten insgesamt 146 Aufgaben. Die Aufgaben stammten aus dem «MINT unterwegs»-Projekt. 24% der Aufgaben wurden doppelt geratet (Brückmann & Duit, 2014, S. 190; Döring & Bortz, 2016, S. 575). In der Literatur wird eine 5-stufige Ratingskala oft als intervallskaliert beschrieben (Busker, 2014, S. 275; Döring & Bortz, 2016, S. 244f.; Wirtz & Caspar, 2002, S. 33f.), deshalb wurden die Urteilübereinstimmungen mittels des Intra Class Correlation Coefficient (ICC) geprüft (Wirtz & Caspar, 2002).
8. Faktorielle Validität, Item- und Reliabilitätsanalyse: Mit den 146 eingeschätzten Aufgaben konnte das Messinstrument statistisch geprüft werden. Hierfür wurde die interne Konsistenz der Items eines Faktors mit dem Cronbach-Alpha-Koeffizienten bestimmt (Moosbrugger & Kelava, 2012, S. 130–137). Bei Items desselben Faktors wurde eine hohe positive Korrelation erwartet. Alpha-Werte $> .80$ gelten als hohe Konsistenzkoeffizienten (Döring & Bortz, 2016, S. 290). Um von den empirisch erhobenen Daten auf die latenten Variablen (Faktoren) zu schliessen, wurden innerhalb der Konstrukte explorative Faktorenanalysen durchgeführt (Moosbrugger & Kelava, 2012, S. 226–334).

Ergebnisse

Die Ergebnisse zur Konstruktion des theoriebasierten und empirisch geklärtem Itempools (Teilschritte 1 bis 4) sind zusammengefasst in Abbildung 1 dargestellt.

Abbildung 1. Übersicht der Konstruktion des theoriebasierten und empirisch geklärten Itempools.

Konstrukt (Grundlage: Literatur und Theorien)	1. Konzeptspezifikation: Anzahl Unteraspekt eines Konstrukts (Grundlage: Literatur und Theorien)		2. Operationalisierung (Item-Konstruktion) (Qualität der Items und Konsens-Rating)		3. Experten:innenbefragungen Rating MINT-Aufgabe 4.1/10.1b		4. Experten:innen-Pretest Rating MINT-Aufgabe 4.1/10.1b		Theoriebasiertes und empirisch geklärtes Itempool
	Konzept	Anzahl	Items	Items	Items	Items	Items		
Kompetenzabbild	3	3	9 Items	9 Items umformulieren	ICC(C,1): .71**	9 Items			9 Items
Lebensweltbezug	3	3	15 Items	- 3 Items 3 Items umformulieren	ICC(C,1): .67**	12 Items			12 Items
Lernendenvorstellung	3	3	17 Items	- 5 Items	ICC(C,1): .83**	12 Items			12 Items
Wissensart	4	4	16 Items	- 3 Items 3 Items umformulieren	ICC(C,1): .78**	13 Items			13 Items
Wissensaktivität	4	4	19 Items	- 5 Items 2 Items umformulieren	ICC(C,1): .69**	14 Items			14 Items
Repräsentationsform	3	3	12 Items	- 1 Item	ICC(C,1): .68**	11 Items			11 Items
Offenheit	7	7	20 Items	+ 2 Items 4 Items umformulieren	ICC(C,1): .87**	22 Items			22 Items
Lernunterstützung	3	3	14 Items	- 4 Items	ICC(C,1): .91**	10 Items			10 Items
Lernweg	3	3	14 Items	- 3 Items 1 Items umformulieren	ICC(C,1): .63**	11 Items			11 Items

Signifikanzniveaus: ** $p < .001$; * $p < .01$; . $p < .05$

Basierend auf den theoretischen Grundlagen wurden neun Konstrukte mit unterschiedlicher Anzahl Unteraspekte angenommen. Die Expert:innenbefragungen führte zu einer qualitativen Überarbeitung des Itempools. Der dritte Expert:innen-Pretest zeigt eine Beurteiler-Übereinstimmung ICC(C,1) von .63 bis .91 ($M = 0.75$, $SD = 0.093$). Die ICC Werte können nach Koo und Li (2016) als moderat/durchschnittlich (zwischen .50 bis .75) gut (zwischen .75 bis .9) und exzellent/sehr gut (größer .9) gewertet werden. Am Ende des iterativen Konstruktionsprozesses beinhaltet der theoriebasierte und empirisch geklärte Pool 114 Items.

Zur Überprüfung der Inhaltsvalidität (Teilschritt 5) wurde ein digitales Closed Card Sorting von Fachpersonen durchgeführt. Die 12 Fachdidaktiker:innen teilten die Items den theoretischen Unteraspekten der Konstrukte zu. Es wurde für jedes Item die relative Häufigkeit bzgl. der Zuordnung gegenüber der theoretischen Setzung berechnet. Eine 100% Zuordnung bedeutet, dass alle Fachdidaktiker:innen das Item dem vorgesehenen theoretischen Konstrukt zugeordnet haben. Als Cut-off-Wert diente eine Übereinstimmung, die kleiner als 45% ist. Nach diesem Prüfverfahren beinhaltete der Itempool noch 107 Items mit einer relativen Zustimmung von 81% bis 100% ($M = 94$, $SD = 0.06$) pro Item. Die zwei theoretischen Substrukture Lernendenvorstellung – abgleichen und Wissensaktivität – transferieren mussten leicht angepasst werden. Um die Interrater-Reliabilität und Durchführungsobjektivität (Teilschritt 6 und 7) zu testen, beurteilten zwei geschulte Fachdidaktiker unabhängig voneinander fünf bis sechs zufällig aus jeder MINT-Unterwegs-Box ausgewählte Aufgaben ($N = 35$) mit dem vorläufigen Messinstrument. Die Item-Korrelationen nach Pearson und Spearman ergeben Werte von .5 bis .9 ($M = 0.72$, $SD = 0.10$). Die ICC(A,1) Werte bewegen sich zwischen .49 bis .87 ($M = .69$, $SD = 0.10$). Mittels Promax-rotierten Faktorenanalysen konnte von den empirisch erhobenen Daten auf die latenten Variablen (Faktoren) geschlossen werden (Teilschritt 8) (Moosbrugger & Kelava, 2012, S. 326–334). Durch die Zusammenfassung mehrerer Konstrukte (angelehnt an Luthiger et al., 2018 und Neubrand, 2002) gewinnt das Instrument an Übersichtlichkeit. Die erste Faktorenanalyse beinhaltet die Skalen Kompetenzabbild und Lebensweltbezug mit je drei Subskalen (siehe Tab.

1). Die zweite Faktorenanalyse besteht aus den Skalen Lernendenvorstellung, Wissensart, Wissensaktivität und Repräsentationsform mit einer bis vier Subskalen. Die dritte Faktorenanalyse schliesst die Skalen Offenheit, Lernunterstützung und Lernweg mit je drei Subskalen mit ein. Die Tabelle 1 stellt zusammenfassend die drei Faktorenanalysen mit ihren jeweiligen Teststatistiken und der Gesamtvarianz dar.

Tabelle 1. Teststatistik der Faktorenanalysen, Bartlett-Test, KMO-, MAP Test und Gesamtvarianz

Faktorenanalyse	χ^2	df	p	KMO	MAP	Gesamt-varianz
Kompetenzabbild und Lebensweltbezug	4033.18	153	< .001	.869	6	91%
Lernendenvorstellung, Wissensart, Wissensaktivität und Repräsentationsform	4740.22	630	< .001	.748	11	79%
Offenheit, Lernunterstützung und Lernweg	4967.43	435	< .001	.82	9	85%

Anmerkungen. χ^2 = Chi-Quadrat Bartlett-Test; df = Freiheitsgraden; p = Signifikanzniveau; KMO = Kaiser-Meyer-Olkin Measure of Sampling Adequacy; MAP = Minimum-Average-Partial-Test

Sowohl die Bartlett-Test's als auch das Kaiser-Meyer-Olkin Kriterium weisen auf die Eignung der Items für Faktorenanalysen hin. Aufgrund des Screeplots (nicht dargestellt) und Minimum-Average-Partial-Test (MAP) wurde den empfohlenen Faktorenlösungen gefolgt. Die Reliabilitätsanalyse der einzelnen Skalen ergibt Cronbach's α -Werte zwischen .824 und .988 ($M = 0.93$, $SD = 0.039$). Die Itemschwierigkeit liegt mit einem Mittelwert von 0.5 ($SD = 0.17$) im guten Bereich.

Diskussion und Ausblick

Das empirisch erarbeitete Aufgaben-Analyse-Instrument kann dazu verwendet werden, die Qualität von Aufgaben differenziert nach neuen Skalen zu analysieren und in Form von Aufgabenprofilen darzustellen.

Einschränkend muss erwähnt werden, dass eine Überprüfung der Faktorstruktur der Skalen mittels konfirmatorischer Faktorenanalysen bzw. Strukturgleichungsmodell aussteht. Es ist zu ergänzen, dass die Subskala Lernendenvorstellung – abgleichen, aufgrund der Validierung, ausschliesslich zwei Items beinhaltet. Eine Erweiterung auf drei bis vier Items bietet sich entsprechend an.

Die Mittelwerte der Fachdidaktiker Intra Class Correlation Coefficient sind nach Koo und Li (2016) im moderaten Bereich. Um die Gütemasse der Übereinstimmungen zu erhöhen, bräuchte es weitere Schulungen und Analysen von Aufgaben. Auch wären bei gewissen Skalen (z.B. Subskala Lebensweltbezug – konstruiert) die Einschätzungen der Lernenden sehr interessant. Die Beurteilung durch Lernende wird in zukünftigen Vorhaben aufgenommen, geht es bei der Kategorie Lebensweltbezug im Wesentlichen um ihre Erfahrungswelt. Auch Lehrpersonen müssen bei zukünftigen Forschungsprojekten in den Fokus genommen werden. So stellte sich beispielsweise bei den Untersuchungen von Blömeke et al. (2006) zu Mathematikaufgaben heraus, dass Aufgaben in der Potenzialeinschätzung keine sozialen Interaktionen vorsahen, die Lehrperson jedoch das abschliessende Reflexionsgespräch sehr stark lenkte und somit die Umsetzung im Unterricht von der Intention abwich. Dies deutet darauf hin, dass ein gelegentlicher Abgleich des Aufgabenpotentials mit dem effektiven Unterricht gewinnbringend sein kann. Des Weiteren sind Aufgabenanalysen in anderen Fachdisziplinen/-didaktiken anzustreben, um das Instrument in diesen zu testen und nutzbar zu machen.

In einem nächsten, noch anstehenden Forschungsschritt, werden die vorliegenden Aufgaben mittels Clusteranalysen nach ihren didaktischen Funktionen gruppiert (Beispielsweise nach

Konfrontations- und Übungsaufgaben), mit den theoretischen Aufgabentypen des LUKAS-Modells verglichen und die Ergebnisse diskutiert. Mit diesem Vorhaben wird untersucht, ob sich die Aufgabentypen nicht nur theoretisch, wie beispielsweise Luthiger et al. (2018) vorschlagen aufteilen lassen, sondern auch datenbasiert.

Referenzen

- Adamina, M. (2010). Mit Lernaufgaben grundlegende Kompetenzen fördern. In P. Labudde (Hrsg.), *Fachdidaktik Naturwissenschaft: 1. - 9. Schuljahr* (1. Aufl.). Haupt.
- Blömeke, S., Risse, J., Müller, C., Eichler, D., & Schulz, W. (2006). Analyse der Qualität von Aufgaben aus didaktischer und fachlicher Sicht. Ein allgemeines Modell und seine exemplarische Umsetzung im Unterrichtsfach Mathematik. *Unterrichtswissenschaft*, 34(4), 330–357.
- Bölsterli Bardy, K., & Wilhelm, M. (2018). Von kompetenzorientierten zu kompetenzfördernden Aufgaben im Schulbuch. *Erziehung & Unterricht*, 168(1/2), 121–129.
- Brückmann, M., & Duit, R. (2014). Videobasierte Analyse unterrichtlicher Sachstrukturen. In D. Krüger, I. Parchmann, & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 189–201). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37827-0_16
- Büchter, A., & Leuders, T. (2006). Ein Aufgabenmodell für die Praxis. Einschätzung, Auswahl und Entwicklung von Mathematikaufgaben. *Praxis der Naturwissenschaften - Chemie in der Schule*, 55, 16–20.
- Büchter, A., & Leuders, T. (2016). *Mathematikaufgaben selbst entwickeln: Lernen fördern - Leistung überprüfen* (7., überarbeitete Neuauflage). Cornelsen.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte und erw. Aufl., [Nachdr.]). Pearson Studium.
- Busker, M. (2014). Entwicklung eines Fragebogens zur Untersuchung des Fachinteresses. In D. Krüger, I. Parchmann, & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 269–281). Springer Spektrum.
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage). Springer.
- Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., Löwen, K., Brunner, M., & Kunter, M. (2006). *Klassifikationsschema für Mathematikaufgaben. Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt*. Max-Planck-Institut für Bildungsforschung.
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kuckartz, U. (2018). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (4. Auflage). Beltz Juventa.
- Luthiger, H., Wilhelm, M., Wespi, C., & Wildhirt, S. (Hrsg.). (2018). *Kompetenzförderung mit Aufgabensets: Theorie - Konzept - Praxis* (1. Auflage). hep, der bildungsverlag.
- Maier, U., Bohl, T., Kleinknecht, M., & Metz, K. (2014). Allgemeine Didaktik und ein Kategoriensystem der überfachlichen Aufgabenanalyse. In P. Blumschein (Hrsg.), *Lernaufgaben – Didaktische Forschungsperspektiven* (S. 35–51). Verlag Julius Klinkhardt.
- Maier, U., Kleinknecht, M., Metz, K., Schymala, M., & Bohl, T. (2010). *Entwicklung und Erprobung eines Kategoriensystems für die fächerübergreifende Aufgabenanalyse*. Univ., Lehrstuhl für Schulpädagogik.

- Moosbrugger, H., & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (2., aktualisierte und überarbeitete Auflage). Springer.
- Müller, A., & Helmke, A. (2008). Qualität von Aufgaben als Merkmale der Unterrichtsqualität – verdeutlicht am Fach Physik. In J. Thonhauser (Hrsg.), *Aufgaben als Katalysatoren von Lernprozessen: Eine zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, allgemeiner Didaktik und Fachdidaktik* (S. 31–46). Waxmann.
- Mummendey, H. D., & Grau, I. (2014). *Die Fragebogen-Methode: Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung*. Hogrefe Verlag.
- Neubrand, J. (2002). *Eine Klassifikation mathematischer Aufgaben zur Analyse von Unterrichtssituationen. Selbsttätiges Arbeiten in Schülerarbeitsphasen in den Stunden der TIMSS-Video-Studie*. Franzbecker.
- Reinfried, S. (2016). *Kompetenzorientierte Lernaufgaben – mehr als alter Wein in neuen Schläuchen?* *Geographie aktuell & Schule*, 38/223, 4-14. http://www.sibylle-reinfried.ch/files_public/Reinfried_2016_Lernaufgaben.pdf
- Reusser, K. (2014). *Aufgaben – Träger von Lerngelegenheiten und Lernprozesse im kompetenzorientierten Unterricht*. Seminar, 4/2014, 77-101.
- Schilb, S. (2005). Card Sorting – Techniken im Überblick. *i-com*, 4(1/2005). <https://doi.org/10.1524/icom.4.1.49.61671>
- Wandke, H., Kain, S., & Struve, D. (2009). *Mensch & Computer 2009: 9. fachübergreifende Konferenz für interaktive und kooperative Medien - Grenzenlos frei*. Oldenbourg Verlag. <https://doi.org/10.1524/9783486598551>
- Wilhelm, M., Luthiger, H., & Wespi, C. (2014). *Kategoriensystem für ein kompetenzorientiertes Aufgabenset*. Luzern: Entwicklungsschwerpunkt Kompetenzorientierter Unterricht, Pädagogische Hochschule Luzern.
- Winkler, I. (2011). *Aufgabenpräferenzen für den Literaturunterricht: Eine Erhebung unter Deutschlehrkräften*. VS Verlag für Sozialwissenschaften / Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-531-92698-8>
- Wirtz, M. A., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe, Verl. für Psychologie.