Problemtypenbasierte Kompetenzmodellierung beim praktisch-naturwissenschaftlichen Arbeiten

Design, Validierung und Einsatz von Aufgaben zum effektbasierten Vergleichen

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Pitt Hild

Zürich, 2020

Erste Gutachterin: Prof. Dr. Ilka Parchmann
Zweite Gutachterin: Prof. Dr. Susanne Metzger

Tag der mündlichen Prüfung: 02. Dezember 2019
Zum Druck genehmigt: 02. Dezember 2019

gez. Prof. Dr. Frank Kempken, Dekan

Zusammenfassung

In der vorliegenden Promotionsarbeit standen die Entwicklung, die Validierung sowie der Einsatz von Aufgaben zum praktisch-naturwissenschaftlichen Arbeiten im Unterricht der Sekundarstufe I im Vordergrund. In einem ersten Schritt wurden, basierend auf einem problemtypenbasierten Kompetenzstufenmodell, exemplarisch für 2 Ausprägungen des praktisch-naturwissenschaftlichen Arbeitens (dem *Vergleichen* und dem *Beobachten*), zu fördernde Fähigkeiten und Fertigkeiten der Schülerinnen und Schüler beschrieben. Eine durch die Kompetenzmodellierung gesetzte a priori Stufung der zu erreichenden (experimentellen) Kompetenzen (Kapitel 2), konnte mittels neuer hands-on Testaufgaben für alle verwendeten Aufgabenkontexte validiert werden. Nebst Hinweisen zur strukturellen Validität der Kompetenzmodellierung (Kapitel 6), wurden weitere Aspekte einer gelungenen Validierung (externe Validität und Generalisierbarkeit des Messinstruments) untersucht (Kapitel 7). Mit den zur Validierung des Messmodells genutzten Leistungs- bzw. Testaufgaben konnten in einem weiteren Teil Lernaufgaben entwickelt werden (Kapitel 8) und somit eine gewisse Anschlussfähigkeit und Nützlichkeit des Kompetenzmodells für die Unterrichtsgestaltung und Inidivualdiagnostik gewährleistet werden. Abschlieβend wurden die experimentellen Aufgaben in einer Interventionsstudie mit Jugendlichen aus leistungsschwachen Klassen der Jahrgangsstufe 7 eingesetzt, mit dem Ziel durch gezieltes kompetenzbezogenes Feedback fehlende Kompetenzen im Bereich des praktisch-naturwissenschaftlichen Arbeitens zu erwerben und weiterzuentwickeln.

Die Ergebnisse zeigen keine statistisch signifikanten Unterschiede in der Entwicklung der experimentellen Kompetenz zwischen Schülerinnen und Schülern der Interventionsgruppen bzw. den einzelnen Interventionsgruppen und der Kontrollgruppe. Das zentrale Problem für die statistisch nicht signifikanten Ergebnisse scheint die zu geringe Größe der Stichprobe zu sein. Insbesondere in einem Falle deutet die Effektstärke des Unterschieds zur Kontrollgruppe darauf hin, dass hier Potenzial vorhanden sein könnte, um die experimentelle Kompetenz bei Schülerinnen und Schüler aus leistungsschwachen Klassen der Sekundarstufe I zu fördern. Optimierungsmöglichkeiten mit Blick auf allfällige neue Studien werden diskutiert.

Abstract

In this dissertation, the design, validation, and practicality of hands-on tasks used to evaluate and foster students' expertise in 'doing science' were focussed. In a first step, performance tasks were developed that were based on an already published competence-based progression model. In this model, skills and abilities (mostly strategies) students of lower secondary school need to have when *doing comparative investigations* or *making observations* in science lessons, are described as different standards to be achieved. An a priori learning hierarchy along the different standards (chapter 2) could be validated using the performance tasks. This hierarchy is independent of the tasks' context. Next to evidence found for structural validity of the competence-based model, described in chapter 6, further aspects of validity (external validity and generalisability) were examined in chapter 7. In a next step, learning tasks based on the validated hierarchy and the performance tasks used in the assessments were developed (chapter 8) to insure a certain practicality and utility of the competence-based model (task design and monitoring) in daily school lessons. Finally, the hands-on tasks were used in an intervention study with 7 grade students from low-achievement levels using competence-based feedback (formative assessment). The aim of this study was to evaluate and develop skills and abilities when doing science. The results show no statistically significant differences between students from the experimental groups and the control group. The central problem for the statistically not significant results seems to be the small sample size. In one case in particular, the effect size indicates that there might be potential here to foster the expertise in ´doing science´ among secondary school students from low-achievement levels. Means for optimization regarding possible new studies are discussed.

Inhaltsverzeichnis

1 Einleitung und Zielsetzung

1.1 Einleitung

Praktisch-naturwissenschaftliches Arbeiten ist ein Grundpfeiler der Naturwissenschaften und ist somit auch im naturwissenschaftlichen Unterricht unverzichtbar. Dies spiegeln auch nationale Bildungsstandards und Lehrpläne wider (z. B. Erkenntnisgewinnung in Deutschland, siehe KMK, 2004 oder Wesen und Bedeutung von Naturwissenschaften und Technik verstehen in der Schweiz, siehe D-EDK, 2015). Unter anderem sollen Schülerinnen und Schüler am Ende der obligatorischen Schulzeit unterschiedliche Denk- und Arbeitsweisen in den Naturwissenschaften nachvollziehen können (Sommer, Wambach-Laicher & Pfeifer, 2018; Vorholzer, 2016). Lehrpersonen sind also aufgefordert, möglichst authentische Lernsituationen zu schaffen, welche das praktisch-naturwissenschaftliche Arbeiten ermöglichen und fördern (Hofstein et al., 2019). Im Unterricht erfolgt dies vor allem durch den Einsatz experimenteller Lernarrangements. Im deutschen Physikunterricht zum Beispiel steht etwa 2/3 der Unterrichtszeit im Zusammenhang mit dem Einsatz von Experimenten (Börlin, 2012; Schreiber & Theyßen, 2019; Tesch & Duit, 2004).

Experimentelle Lernarrangements fokussieren dabei *experimentelle Kompetenzen* von Schülerinnen und Schülern und erfüllen mehrere Funktionen: Schülerinnen und Schüler sollen

- *das Experimentieren* als möglichen Weg zur Erkenntnisgewinnung kennenlernen (im Sinne von *scientific inquiry*),
- befähigt werden, Probleme zu lösen,
- Experimentierabläufe kennen (und die jeweils "richtige" Methode auswählen),
- das "Handwerkszeug" zum praktisch-naturwissenschaftlichen Arbeiten erlernen (welche Messgeräte muss ich einsetzen, wie benutze ich sie, …),
- nicht zuletzt dank geeigneter Lernarrangements auch über ihren jeweiligen Kompetenzstand informiert und in ihren *experimentellen Kompetenzen* gefördert werden.

Gerade im letzten Bereich (Diagnostik/Förderung) fühlen sich vor allem angehende Lehrpersonen nicht sehr kompetent (z. B. Klieme & Warwas, 2011). Die Frage, wie (und auch mit welchen Hilfsmitteln) Lehrpersonen individuelle Lernunterstützungen zum praktisch-naturwissenschaftlichen Arbeiten planen und adaptiv, sprich an den Lernstand und Lernfortschritt der Schülerinnen und Schüler angepasst, einsetzen sollen, ist bis heute nicht ausreichend evidenzbasiert beantwortet (vgl. Krammer, 2009). Die hier vorgestellte Arbeit knüpft an dieses letzte Desideratum an.

1.2 Zielsetzung

Ein erstes Ziel dieser Arbeit war es herauszufinden, wie experimentelle Kompetenzen von Schülerinnen und Schülern beschrieben und diagnostiziert werden können. Das Konstrukt hinter dem Schlagwort der experimentellen Kompetenzen musste zuerst theoretisch eingerahmt und Kompetenzen modelliert werden (Schecker & Parchmann, 2006). Ein großer Teil dieser Denkarbeit fand im Projekt ExKoNawi der Pädagogischen Hochschule Zürich statt (z. B. Gut et al., 2014; Projektnummer SNF 162684). Für die Kompetenzmodellierung wurde der *Problemtypenansatz* gewählt (siehe Details bei Gut & Mayer, 2018, sowie in Abschnitt 2.3).

Im Projekt ExKoNawi wurde zur Beschreibung des praktisch-naturwissenschaftlichen Arbeitens zwischen experimentellen Problemtypen (fragengeleitetes Untersuchen, effektbasiertes Vergleichen, skalenbasiertes Messen, kategoriengeleitetes Beobachten) unterschieden, welche jeweils andere experimentelle Fähigkeiten und Fertigkeiten der Schülerinnen und Schüler in den Fokus stellen. Zu einigen dieser Problemtypen, wie zum Beispiel dem *fragengeleiteten Untersuchen* wurde bereits viel geforscht (z. B. Emden, 2011; Hammann, Phan & Bayrhuber, 2007), zu anderen Problemtypen, wie dem *effektbasierten Vergleichen* lässt sich nur wenig Forschungsliteratur finden (Hungerfold & Miles, 1969; Meyer & Carlisle, 1996; Tomera, 1974).
Ein weiteres Ziel dieser Arbeit war es deshalb, Aufgaben zum effektbasierten Vergleichen zu modellieren und zu validieren, experimentelle Kompetenzen in diesem Bereich formativ zu beurteilen (Umsetzung in der Praxis) und Hinweise auf die erfolgreiche Art der Lernunterstützung zu generieren.

2 Theoretischer Rahmen

2.1 Praktisch-naturwissenschaftliches Arbeiten

Unter praktisch-naturwissenschaftlichem Arbeiten in der Schule werden „…alle beobachtenden und experimentellen Aktivitäten im naturwissenschaftlichen Unterricht…" (Wilhelm & Kunz, 2016, 126) zusammengefasst. Die beiden Autoren Wilhelm und Kunz unterscheiden hier zwischen acht verschiedenen Ausprägungen des praktisch-naturwissenschaftlichen Arbeitens: dem *Betrachten*, dem *Beobachten*, dem *Messen*, dem *Studieren*, dem *Erkunden*, dem *Vergleichen*, dem *Versuchen* und dem *Experimentieren*[1].

Die Ziele des praktisch-naturwissenschaftlichen Arbeitens (im Unterricht) haben sich seit dem Anfang des 20. Jahrhunderts stark verändert. Während Dewey (1910, 192) noch bemängelte, dass z. B. mit dem Einsatz von Experimenten zu häufig allein die manuellen Fähigkeiten gefördert würden als seien sie allein schon Bildungsziel, stehen fachspezifisch manuelle Fertigkeiten wie beispielsweise das Bauen und Kalibrieren eigener Messinstrumente kaum mehr im Fokus des naturwissenschaftlichen Unterrichts. Dafür fallen heutzutage unter das praktisch-naturwissenschaftliche Arbeiten das allgemeine Problemlösen (bereits erwähnt bei Bruner, 1962), sowie Methoden der Erkenntnisgewinnung und allgemeine naturwissenschaftliche Denk- und Arbeitsweisen (Details bei Hofstein & Lunetta, 2004 oder Vorholzer, 2016).

Diese Forderungen an das praktisch-naturwissenschaftliche Arbeiten stellen Lehrpersonen und Lernende vor einige Herausforderungen, mit welchen sich die Naturwissenschafts-didaktik in den letzten Jahrzehnten beschäftigte. So wurden Vorschläge für Modelle von experimentellen Kompetenzen gemacht (z. B. Mayer, Grube & Möller, 2008) sowie beschrieben, wie sich experimentelle Kompetenzen von anderen fachspezifisch naturwissen-schaftlichen Kompetenzen unterscheiden (z. B. Millar et al, 1996) und welche Kompetenzen Schülerinnen und Schüler beim praktisch-naturwissenschaftlichen Arbeiten erwerben sollen (z. B. Börlin, 2012, 27).

[1] *Experimentieren* wird hier als hypothesengeleitetes Forschen durch Manipulation der unabhängigen Variable verstanden und führt im Gegensatz zum konfirmatorischen *Versuchen*, zu neuen Erkenntnissen (Wilhelm & Kunz, 2016, 128).

2.2 Das Experiment als Lerngelegenheit

Mit Experimenten können im Unterricht gleich mehrere Ziele verfolgt werden. Gemäß Hodson (2009) stehen dabei drei Grundziele im Fokus:

1) das Experiment als fachspezifische Arbeitsweise im naturwissenschaftlichen Unterricht (knowing science) um Phänomene kennenzulernen; naturwissenschaftliche Sachverhalte zu verdeutlichen; Fachsprache in konkreten Kontexten eingebettet, zu üben; naturwissenschaftlich nicht adäquate in naturwissenschaftlich adäquate Konzepte umzudeuten; u.s.w.

2) das Experiment als zentraler Bestandteil naturwissenschaftlichen Arbeitens (doing science) um neue Erkenntnisse zu gewinnen, indem u. a. Experimente geplant, Hypothesen überprüft, Daten ausgewertet werden.

3) Reflektieren der Bedingungen und Eigenschaften des Experiments als Methode der Erkenntnisgewinnung (knowing about science) um zu verstehen was eine Hypothese ist; wie richtig beobachtet wird (siehe dazu Kapitel 7); was unter Messunsicherheit verstanden wird; wieso Messungen wiederholt werden; u.s.w.

Zudem erfüllt das Experiment auch unterrichtsmethodische Funktionen. Dank Experimenten entstehen Lerngelegenheiten, welche dazu beitragen können, den Austausch zwischen den Lernenden zu fördern, eine gewisse Rhythmisierung im naturwissenschaftlichen Unterricht zu gewährleisten (Börlin, 2012, 12), das Interesse bei Schülerinnen und Schüler für bestimmte Kontexte zu wecken und weitere lernrelevante Faktoren (wie z. B. authentisches Lernen, Problemlösefähigkeit,…) zu unterstützen (z. B. Blömeke et al., 2006).

2.3 Modellierung experimenteller Kompetenzen

Die unterschiedlichen Ziele, welche mit dem Einsatz von Experimenten verfolgt werden (können), erschweren eine Festlegung auf ein einheitliches Konstrukt hinter dem Begriff der experimentellen Kompetenzen. Letztere umfassen „…das Wissen und die Fähigkeit, durch gezielte handelnde Auseinandersetzung mit der Natur Daten zu gewinnen, diese vor dem Hintergrund von Modellen und Theorien zu interpretieren und dadurch Wissen und Erkenntnisse über die Natur abzuleiten" (Gut & Mayer, 2018, 122). Experimentelle Kompetenzen können dabei jedoch nicht nur situationsgebundene fachspezifisch inhaltliche (z. B. Verständnis des Temperaturkonzeptes) und methodische (z. B. Temperaturmessung), sondern auch überfachlich methodische Fertigkeiten und Fähigkeiten (z. B. exaktes und sorgfältiges Arbeiten), sowie Dispositionen (z. B. Selbstkonzept, aktuelle Lernmotivation) umfassen (z. B. Blömeke, Gustafsson & Shavelson, 2015).

Hinzu kommt, dass Schülerinnen und Schülern ihre Kompetenzen, auch im Bereich des praktisch-naturwissenschaftlichen Arbeitens, grundsätzlich nach mehreren Progressionslogiken weiterentwickeln können. Zum Beispiel werden Schülerinnen und Schüler "kompetenter", indem sie lernen, zunehmend komplexere Probleme zu lösen, Fragen und Probleme qualitativ vertiefter, differenzierter, eigenständiger, in zunehmend anspruchsvolleren fachlichen Kontexten und mit entsprechenden Informationsmitteln und medialen Repräsentationen zu lösen (vgl. Adamina & Hild, 2019; Gut et al., 2014).

Eine Kompetenzmodellierung für das praktisch-naturwissenschaftliche Arbeiten verlangt deshalb im Vorfeld eine übergeordnete theoretische Rahmung: Je nach Forschungsfrage, aber auch je nach Situation im Unterricht, müssen im Vorfeld Abgrenzungen getroffen sowie festgelegt werden, welche Teilkompetenzen zu den experimentellen Kompetenzen zählen sollen und welche nicht. Gut & Mayer (2018) sprechen hier von:

i) den psychologischen Grundlagen experimenteller Kompetenzen (im Vorfeld muss u. a. festgelegt werden, welche Wissensarten und Denkprozesse dazu zählen),

ii) der Spezifizierung des Konstrukts (fachliche Kontexte müssen eingegrenzt werden, Handlungen und Dispositionen, die erwartet werden, müssen beschrieben werden),

iii) einer inneren Differenzierung der experimentellen Kompetenzen. Unter anderem soll die Frage beantwortet werden, ob die Erkenntnisgewinnung in Teilschritte zerlegt (Teilprozessansatz, z. B. Emden, 2011) oder Experimentieren als integraler Problemlöseprozess (Problemtypenansatz, vgl. Ruiz-Primo & Shavelson, 1996; Shavelson, 1992) angesehen werden soll.

In den letzten Jahren wurden in der deutschsprachigen naturwissenschaftsdidaktischen Forschung gerade für den Kompetenzbereich der naturwissenschaftlichen Erkenntnisgewinnung unterschiedliche Struktur- und Progressionsmodelle postuliert und zum Teil validiert, die sich mit dem praktisch-naturwissenschaftlichen Arbeiten (Mayer, Grube & Möller, 2008; Metzger et al. 2014a) oder spezifischer mit dem Experimentieren befassen (Emden, 2011; Hammann, Phan & Bayrhuber, 2007; Kauertz et al., 2010; Neumann et al., 2007). In diesen Modellen werden der Kompetenzaufbau resp. die Kompetenzentwicklungen nach verschiedenen Stufen des Wissens und Könnens konstruiert. Für die Aufgabenkonstruktion bilden solche Modellierungen eine wichtige Grundlage.

Im Rahmen dieser Dissertation wurden, wie im Projekt ExKoNawi, experimentelle Kompetenzen mit dem Problemtypenansatz modelliert. In den folgenden Abschnitten wird auf diese Modellierung genauer eingegangen.

2.4 Problemtypenbasierte Kompetenzmodellierung im Projekt ExKoNawi

Das Projekt ExKoNawi, *Experimentelle Kompetenzen in den Naturwissenschaften* knüpfte seit 2012 an die Ergebnisse unterschiedlicher large-scale Experimentiertests mit dem Ziel an, die Validitätsprobleme (lokale Itemabhängigkeiten, kompetenzirrelevante Anforderungen, Generalisierbarkeit und Interpretierbarkeit der hands-on Testaufgaben) in zufriedenstellender Weise zu lösen (siehe Details bei Gut, 2012; Konsortium HarmoS Naturwissenschaften+, 2010; Ramseier, Labudde & Adamina, 2011).

Aufgrund des integrierten naturwissenschaftlichen Unterrichts in der Schweizer Volksschule wird für die Messung experimenteller Kompetenzen ein Modell benötigt, das biologische, chemische und physikalische Aspekte des praktisch-naturwissenschaftlichen Arbeitens berücksichtigt. Basierend auf dem Problemtypenansatz (Gut et al. 2014), wurden für unterschiedliche Experimentiersituationen im Unterricht zuerst experimentelle Problemtypen definiert (siehe Tab. 2.1) und dann a priori für jeden Problemtyp festgelegt, welche (Teil)-kompetenzen, im Folgenden auch *Qualitätsstandards* (QS) genannt, besonders im Fokus stehen (Gut et al. 2014).

Tab. 2.1 Fachspezifische Kompetenzbeschreibungen unterschiedlicher Problemtypen.

Problemtypen	Kategoriengeleitetes Beobachten	*Phänomene anhand gegebener Fragen (Kategorie) beschreiben und vergleichen*
	Skalenbasiertes Messen	*Quantitative Größen mit gegebenen Messinstrumenten (Skala) genau messen*
	Fragengeleitetes Untersuchen	*Korrelative Zusammenhänge zwischen gegebenen Variablen (Frage) untersuchen*
	Effektbasiertes Vergleichen	*Objekte anhand einer gegebenen Eigenschaft (Effekt) experimentell vergleichen*
	…	…

Zum Kompetenzmodell aus dem Projekt ExKoNawi wurden in einer ersten Modellierung folgende vier Problemtypen genauer untersucht:

- *kategoriengeleitetes Beobachten*,
- *skalenbasiertes Messen*,
- *fragengeleitetes Untersuchen* sowie
- *effektbasiertes Vergleichen.*

Bei jeder Modellierung wurde von einer Lernhierarchie ausgegangen: Schülerinnen und Schüler erreichen einfacher bzw. häufiger einen ersten als einen zweiten QS, einfacher bzw. häufiger einen zweiten als einen dritten, etc. Die a priori festgelegten Lernhierarchien wurden während der Validierungsstudien (Pilottests 1 bis 3) im Projekt ExKoNawi für alle vier Problemtypen sowie im Rahmen dieser Dissertation speziell für den Problemtyp effektbasiertes Vergleichen mit unterschiedlichen Stichproben und unterschiedlichen Kontexten überprüft.

2.4.1 Effektbasiertes Vergleichen

Aufgaben zum experimentellen Problemtyp effektbasiertes Vergleichen werden definiert als Aufgaben, bei denen Objekte anhand einer gegebenen Eigenschaft verglichen werden sollen und alle anderen Variablen konstant gehalten werden (Solano-Flores & Shavelson, 1997). Für den Problemtyp effektbasiertes Vergleichen, wurden vier QS festgelegt (Abb. 2.1, Abb. 5.1 sowie Fig 6.1).

Die Schülerinnen und Schüler können …

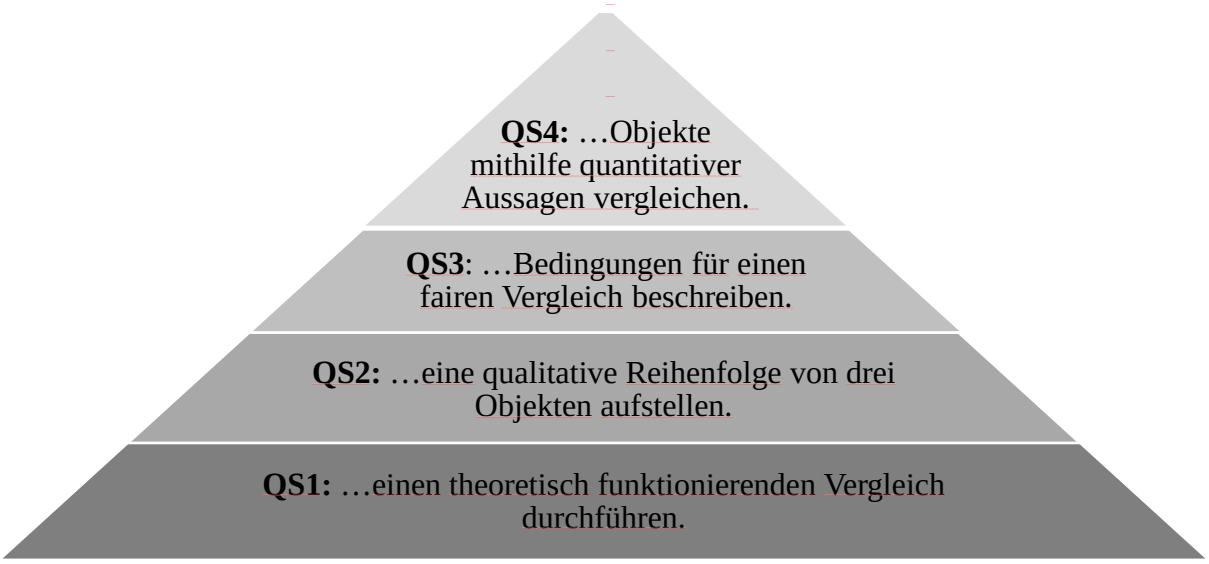


Abb. 2.1 Die vier QS beim effektbasierten Vergleichen – ein Kompetenzstufenmodell.

Beim ersten und zweiten Qualitätstandard wird überprüft ob der vorgeschlagene Vergleich sich theoretisch eignet (für zwei bzw. drei Objekte), um eine gewisse Objekteigenschaft zu überprüfen. Es müssen jedoch keine Daten vorhanden sein und der Vergleich kann auch ohne Variablenkontrolle durchgeführt worden sein. QS 3 wird nur dann erreicht, wenn beim Vergleich explizit erkenntlich wird, dass gewisse Bedingungen eingehalten wurden (Variablenkontrolle). Das letzte QS kann nur erreicht werden, wenn quantitative Daten vorhanden sind, welche Aussagen über die Ausprägung oder Differenzen der Ausprägungen der Objekteigenschaften zulassen.

2.4.2 Kategoriengeleitetes Beobachten

In Anlehnung an Arnold, Wellnitz & Mayer (2010) werden experimentelle Kompetenzen beim "naturwissenschaftlichen" Beobachten durch Fähigkeiten und Fertigkeiten beschrieben, mit denen naturwissenschaftliche Fragen durch das Beobachten von Merkmalen, Phänomenen und Prozessen beantwortet werden können. Auch im Projekt ExKoNawi fallen unter den experimentellen Problemtyp kategoriengeleitetes Beobachten Aufgaben, in denen naturwissenschaftliche Phänomene anhand gegebener Fragen (bzw. Kategorien) beschrieben und verglichen werden (Metzger et al., 2014a; Metzger et al. 2014b).
Für den Problemtyp kategoriengeleitetes Beobachten wurden anfangs fünf QS festgelegt. Jedoch konnte bei den eingesetzten Testaufgaben mittels ersten Validierungsstudien eine Lernhierarchie entlang lediglich dreier Qualitätsstandards aufgedeckt werden (Abb. 2.2).

Die Schülerinnen und Schüler können …

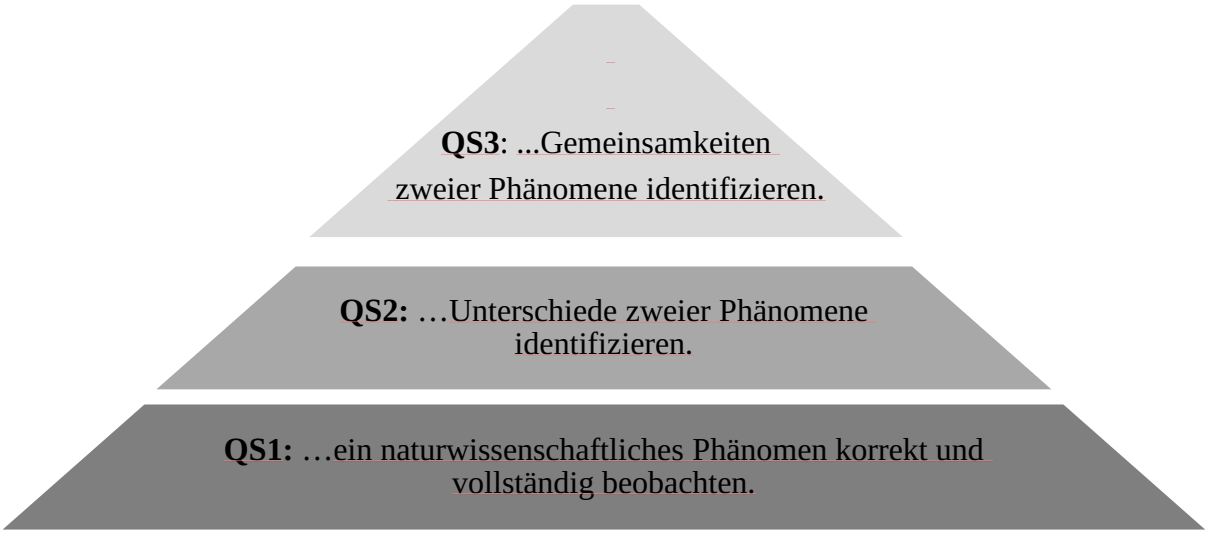


Abb. 2.2 Die drei QS beim kategoriengeleiteten Beobachten – ein Kompetenzstufenmodell.

2.5 Konstruktion von Aufgaben zum effektbasierten Vergleichen

Zu jedem Problemtyp wurden kontextspezifische Testaufgaben und standardisierte Kodiermanuale entwickelt. Insgesamt wurden im Projekt ExKoNawi in drei Schritten 24 hands-on Testaufgaben entwickelt (6 Aufgaben pro Problemtyp), die jeweils in drei disjunkten Tests pilotiert wurden (Gut et al., 2017). Alle Aufgaben mit chemischen Kontexten wurden vom Autor selbst entwickelt (insgesamt acht Testaufgaben). Zum experimentellen Problemtyp effektbasiertes Vergleichen wurden die Testaufgaben in den Pilotierungsphasen 2 und 3 erprobt (graue Schraffierung in Tab. 2.2). Auf deren Validierung wird vor allem in Kapitel 6 eingegangen. Im Rahmen der Dissertation wurden neben den eingesetzten Testaufgaben aus dem Projekt ExKoNawi noch 3 weitere Aufgaben mit chemischen Kontexten zum effektbasierten Vergleichen entwickelt.

Tab. 2.2 Hands-on Testaufgaben zum Problemtyp effektbasiertes Vergleichen mit chemischen Kontexten.

Testaufgabe	**Problem**
Milchprodukte	*Otto und Olivia wollen herausfinden, welches Milchprodukt (Magerquark, 2 unterschiedlich fette Streichkäse) am wässrigsten ist.*
Crèmes	*Sonja und Silvano wollen herausfinden, welche Crème (Körperlotion, Handcrème, Nivea®) am wässrigsten ist.*
Tee	*Claudio und Catherine wollen herausfinden, welcher Tee (Schwarztee, Rooibos- und Früchteinfusion) am sauersten ist.*
Pulver	*Natalia und Nino wollen herausfinden, welches Pulver (Einfach- & Zweifachzucker, Kochsalz rein) am besten in Wasser löslich ist.*
Nüsse	*Xaver und Xenia wollen herausfinden, welche Nuss (Walnuss, Erdnuss, Mandel) am meisten Fett enthält.*

In diesem Abschnitt wird nur auf die Konstruktion von Testaufgaben eingegangen. Falls Lernaufgaben konstruiert werden sollen, sollten weitere lernrelevante Merkmale (wie z. B. Anforderungsmerkmale, Differenzieriungsmöglichkeiten, Unterstützungsangebote) beachtet werden (Details u. a. bei Adamina & Hild, 2019; Luthiger, Wilhelm & Wespi, 2014; Maier, Kleinknecht & Metz, 2010). Diese Merkmale sind jedoch häufig für die Konstruktion von Testaufgaben (zu Validierungszwecken) irrelevant und können sogar dazu führen, dass allgemeingültige Aussagen verhindert werden.

Für die Konstruktion der Testaufgaben wurde ein stark offener Aufgabenstamm gewählt. Der Öffnungsgrad der Aufgaben bezieht sich hier auf den Bewertungsraster in Tabelle 2.3, der aufzeigt, welche Teilbereiche bei der Aufgabenstellung vorgegeben und welche offen (von den Schülerinnen und Schülern ausgearbeitet werden müssen) sind. In vorliegendem Fall wurde den Schülerinnen und Schülern jeweils das zu lösende Problem angegeben, sie mussten jedoch selber eine Methode entwickeln, Daten generieren und interpretieren (entspricht Öffnungsgrad 2 in Tabelle 2.3).

Tab. 2.3 Bewertungsraster für hands-on Aufgaben (in Anlehnung an Tabelle 2 und 3 aus Fay et al., 2007)

Öffnungsgrad	Problem/Fragestellung	Vorgehen/Methode	Lösung
0 geschlossen	***vorgegeben***	***vorgegeben***	***vorgegeben***
	Problem, Vorgehen und Lösung werden den Schülerinnen und Schülern ausgeteilt. Die Schülerinnen und Schülern führen die hands-on Aufgabe durch und überprüfen/vergleichen ihre Lösung mit der Problemlösung.		
1 leicht offen	***vorgegeben***	***vorgegeben***	***offen***
	Problem und Vorgehen werden den Schülerinnen und Schülern ausgeteilt. Die Schülerinnen und Schüler führen die hands-on Aufgabe durch, sammeln und interpretieren Daten, schlagen eine Lösung vor.		
2 stark offen	***vorgegeben***	***offen***	***offen***
	Das Problem wird den Schülerinnen und Schülern ausgeteilt. Die Schülerinnen und Schüler überlegen sich ein Vorgehen, entscheiden welche Daten gesammelt und interpretiert werden, schlagen eine Lösung vor.		
3 ganz offen	***offen***	***offen***	***offen***
	Den Schülerinnen und Schülern wird ein Phänomen gezeigt. Sie entwickeln selber eine Problem- resp. Fragestellung, überlegen sich ein Vorgehen, entscheiden welche Daten gesammelt und interpretiert werden, schlagen eine Lösung vor.		

Die grundlegende Anforderung an die zu entwickelnden Aufgaben war, dass mit ihnen die a priori festgelegten Qualitätsstandards (siehe Abb. 2.1) überprüft werden können.

Entsprechend der Qualitätsstandards lag der Hauptfokus der hands-on Testaufgaben vor allem auf fachspezifisch methodischen (fachmethodischen), nicht auf inhaltlichen Kompetenzen. So wurde z. B. überprüft, ob Schülerinnen und Schüler faire Vergleiche durchführen können. Durch die Fokussierung auf fachmethodische Kompetenzen, mussten die Testaufgaben so konzipiert werden, dass der Einfluss des fachinhaltlichen Vorwissens auf die Testleistung gering ausfällt (siehe dazu auch Bonetti, Gut, Metzger & Walpuski, 2019). Um dies zu gewährleisten, wurden deshalb Kontexte gewählt, welche kaum Inhalt des regulären naturwissenschaftlichen Unterrichts sind (siehe Tab. 2.2). Des Weiteren wurden für diese Testaufgaben Vignetten konzipiert, die den Schülerinnen und Schülern die Problemstellung der jeweiligen Testaufgabe verdeutlichen und dazu führen sollten, dass alle Schülerinnen und Schüler den einzelnen Testaufgaben mit dem möglichst gleichen Wissensstand begegnen sollten (Details zum Einsatz von Vignetten bei von Aufschnaiter, Selter & Michaelis, 2017).

Tabelle 5.3 zeigt den Aufgabenstamm mit den Vignetten (graue Schraffierung). Dieser Stamm wurde im Rahmen der Dissertation bei allen Testaufgaben zum Problemtyp effektbasiertes Vergleichen nicht verändert. Zwei Testaufgabenbeispiele (in komprimierter Form) befinden sich in Tabelle 5.5 (*Nüsse*) und im Anhang (englische Version von *Milchprodukte* aus der Publikation 2). Die Schülerinnen und Schüler hatten für jede hands-on Testaufgabe jeweils 18 Minuten Zeit. In einem 6-seitigen Testheft wurden sie aufgefordert, ihr Vorgehen zu protokollieren und reflektieren.

2.6 Validierung der Aufgaben zum effektbasierten Vergleichen

Zunächst wurden alle Aufgaben in Vorvalidierungsstudien (teils im Projekt ExKoNawi, teils im Rahmen der Dissertation) einzeln mit einer kleinen Stichprobe vorpilotiert (Gut et al., 2017). Im Anschluss an die Bearbeitung wurde jeweils mit den Schülerinnen und Schülern besprochen, inwiefern die Fragestellung (bzw. das Problem und die einzelnen Teilaufgaben, siehe Tab 5.3) verständlich formuliert war.

Für die Beurteilung der experimentellen Kompetenzen sowie zur Überprüfung der Aufgabenvalidität wurden standardisierte Kodiermanuale entwickelt, welche sich für alle Testaufgaben zum experimentellen Problemtyp effektbasiertes Vergleichen aus zehn Kriterien (bezogen auf die vier Qualitätsstandards) zusammensetzten (siehe dazu auch Abschnitt 5.5.1 sowie tab. 6.7).

In zwei Validierungsstudien (Pilottest 2 und Pilottest 3) wurden im Projekt ExKoNawi Testaufgaben zum Problemtyp effektbasiertes Vergleichen untersucht. Die vorgestellte Arbeit baute auf den Ergebnissen dieser beiden Pilotstudien auf.

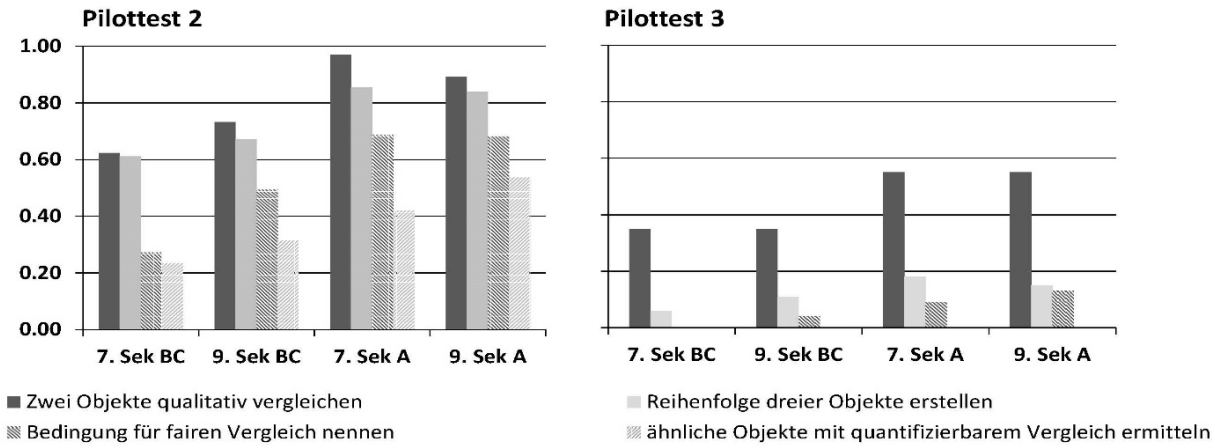


Abb. 2.3 Vergleich der Häufigkeiten der erreichten QS über alle Aufgaben zum effektbasierten Vergleichen bei den Pilottests 2 und 3 (aus Gut et al., 2017).

Abbildung 2.3 zeigt, dass Schülerinnen und Schüler in beiden Pilottests häufiger den ersten als den zweiten und häufiger den zweiten als den dritten QS erreichen. Der vierte QS wurde in Pilottest 3 nie erreicht. Pilottest 3 war anscheinend zu schwierig für die getestete Stichprobe (Gut et al., 2017). Außerdem könnte das schlechte Abschneiden auch auf eine strengere Kodierung und “schwierigere” Fachkontexte zurückzuführen sein. Im Rahmen der

Dissertation wurden die standardisierten Kodiermanuale jedoch nicht mehr verändert. Aber die Aufgaben wurden in den verschiedenen Studien, welche in den folgenden Kapiteln vorgestellt werden, mit jeweils unterschiedlichen Stichproben sowie unterschiedlichen Kodierenden durchgeführt.

Während bei der Auswertung der Pilottests (Projekt ExKoNawi) vor allem die Argumente einer strukturellen Validität der Testaufgaben und des Modells im Fokus standen, wurden die Aufgaben zum effektbasierten Vergleichen sowie das dazugehörige Modell im Rahmen der Dissertation zusätzlich auf weitere Argumente für Validität (externe Validität und Verallgemeinerbarkeit) nach Messick (1996) überprüft (siehe graue Schraffierung in Tab. 2.4). Dabei wurden die Studien um ein weiteres Erhebungsverfahren (Schülerlabor), weitere Stichproben, weitere Kontexte, sowie weitere Messverfahren (Videos und Interviews) ergänzt.

Tab. 2.4 Sechs Validitätsaspekte für die Konstruktion und Validierung von Testaufgaben (nach Messick, 1996; deutsche Version und Beschreibung der Aspekte bei Leuders, 2014).

Aspekt der…	**Beschreibung**
inhaltlichen Validität (content),	*Curriculare und theoretische Absicherung des modellierten Bereichs.*
kognitiven Validität (substantive),	*Passung der kognitiven Prozesse bei der Kompetenzerfassung zum postulierten theoretischen Kompetenzmodell.*
strukturellen Validität (structural),	*Passung von theoretischem Kompetenzmodell und gewähltem psychometrischem Messmodell.*
Verallgemeinerbarkeit (generalisability),	*Angemessenheit einer über die Aufgaben- und Personengruppe hinausgehenden Interpretation.*
externen Validität (external),	*Angemessenheit mit Blick auf konvergente, diskriminante und prädiktive Zusammenhänge mit anderen Konstrukten.*
konsequentiellen Validität (consequential).	*Angemessenheit der Nutzung im pädagogischen oder bildungspolitischen Kontext.*

2.7 Formative Beurteilung experimenteller Kompetenzen

Abbildung 2.3 zeigt deutlich, dass viele Schülerinnen und Schüler die geforderten QS beim erstmaligen Lösen von Testaufgaben zum effektbasierten Vergleichen nicht erreichen. Das Kompetenzstufenmodell (Abb. 2.1) könnte den Lehrpersonen bei der Diagnose und Planung individueller, adaptiver Lernunterstützungen (Scaffolding-Ansatz) behilflich sein (siehe Details unter 8.3). Als eine mögliche Maßnahme adaptiver Lernunterstützung soll an dieser Stelle das Feedback genauer untersucht werden (Details bei Van de Pol, Volman & Beishuizen, 2010).

Für eine an ein Feedback gekoppelte lernbegleitende Diagnose zur Förderung eines Lernprozesses wird häufig der Begriff der *formativen Beurteilung* verwendet (vgl. ler and. Feedback ist hier als integrativer Bestandteil formativer Beurteilungen zu verstehen (vgl. Sadler, 1989; Sandoval & Reiser, 2004). Formative Beurteilungen sollten beispielsweise zur Optimierung des Unterrichts stattfinden (z. B. Bennett, 2011). Für Bloom (1969) sollen des Weiteren formative Beurteilungen dazu beitragen, Feedback und Verbesserungsvorschläge in den unterschiedlichen Etappen des Lernens zu liefern. Im Gegensatz zu summativen Beurteilungen finden bei formativen Beurteilungen keine Urteile oder Bewertungen statt. Unumstritten ist, dass im Unterricht formative und summative Beurteilungen nötig sind und diese bestmöglich miteinander verknüpft sein sollen (Allal, 2010; Shavelson et al., 2008).
Allgemein zählt Feedback bei "korrektem" Einsatz mit zu den wirkungsvollsten Interventionen zur Förderung von Lern- und Entwicklungsprozessen (Black & William 1998; Müller & Ditton, 2014). Aus diesem Grund wurde für die Interventionsstudie im Rahmen der Dissertation das individuelle Feedback als Fördermaßnahme gewählt. Das Feedback bezog sich in dieser Studie auf das Kompetenzmodell aus dem Projekt ExKoNawi (exemplarisch zum effektbasierten Vergleichen) und wird in Anlehnung an Wollenschläger, Möller und Harms (2012) als *kompetenzielles* oder *kompetenzbezogenes* Feedback angesehen. Bis dato existieren nur wenige Studien, in welchen Schülerinnen und Schüler ihre fachmethodischen Kompetenzen im Bereich des praktisch-naturwissenschaftlichen Arbeitens mittels adaptiven Lernunterstützungen verbessern konnten (vgl. Arnold et al., 2017; Ropohl & Scheuermann, 2018; Wollenschläger et al., 2012). Dementsprechend widmen sich gleich zwei von fünf übergeordneten Fragestellungen dieser Dissertation der kompetenzbezogenen Lernunterstützung beim praktisch-naturwissenschaftlichen Arbeiten.

2.8 Fokus auf leistungsschwache Klassen

Im Schuljahr 2017/2018 befanden sich im Kanton Zürich 38.838 12- bis 15-jährige Schülerinnen und Schüler in der Sekundarstufe I (entspricht der 7ten bis 9ten Jahrgangsstufe in Deutschland), welche in 3 Anforderungsniveaus aufgeteilt wird: Gymnasium (18%), Sekundarstufe A (44.5%) sowie Sekundarstufe B und C (37.5%). Schülerinnen und Schüler aus Sonderschulen werden hier dem letzten Anforderungsniveau zugeteilt (Bildungsdirektion Kanton Zürich, 2018). Weitere 3.909 Schülerinnen und Schüler befanden sich in privaten Schulen und Einrichtungen.

Obwohl Studien deutlich zeigen, dass getrennte Schulsysteme dazu tendieren, die Ungleichheiten im Kompetenzerwerb zwischen Schülerinnen und Schülern am Ende der obligatorischen Schulzeit zu verstärken, ist dies das gängige Modell in den Schweizer Kantonen (Felouzis & Charmillot, 2017). Schülerinnen und Schüler des Gymnasiums und der Sekundarstufe A werden zu statistischen Zwecken der Kategorie *erweiterte Ansprüche*, diejenigen aus der Sekundarstufe B/C der Kategorie *Grundansprüche* zugeteilt (SKBF, 2018). In der Kategorie Grundansprüche befinden sich mehr Jungen (56.5 %) als Mädchen und der Anteil der Migrantinnen und Migranten liegt hier rund doppelt so hoch wie in Schulklassen mit erweiterten Ansprüchen (Felouzis & Charmillot, 2017).

In der Interventionsstudie im Rahmen der Dissertation wurde ausschließlich mit Schülerinnen und Schülern aus der Kategorie Grundansprüche gearbeitet (siehe Kapitel 6). Dies hat zwei Gründe:
Zum einen zeigte sich in den Pilotierungsstudien im Projekt ExKoNawi deutlich, dass bei Schülerinnen und Schüler aus der Kategorie Grundansprüche in allen gemessenen Experimentiersituationen das größte Lernpotenzial bestand (Abb. 2.3). Für Länder mit getrennten Schulsystemen konnte nachgewiesen werden, dass die Schulart einen zusätzlichen Einfluss auf die Entwicklung naturwissenschaftlicher Kompetenzen hat und dass diese für Schülerinnen und Schüler aus der Kategorie Grundansprüche (in Deutschland Schülerinnen und Schüler aus den integrierten Haupt-/Realschulen und Realschulen) am geringsten ausfällt (Ivanov, 2011). Zudem belegen die PISA Studien, dass sich die *naturwissenschaftliche PISA Kompetenz* von der 9. zur 10. Klasse am Gymnasium verbesserte, jedoch in den nicht gymnasialen Schularten abnahm (Schiepe-Tiska et al., 2017, 172).

Zum anderen wurden bis dato in fast allen Interventionsstudien, die den Erwerb und Zuwachs experimenteller (inhaltlicher wie methodischer) Kompetenzen durch Lernunterstützungsangebote (z. B. durch *Scaffolding wie Lösungsbeispiele, Strukturierungshilfen, Feedback, Prompts*) untersuchten, ausschließlich mit Schulklassen aus dem Gymnasium oder einzelnen Schülerinnen und Schülern aus beiden Kategorien, die freiwillig an Studien teilnahmen, gearbeitet (vgl. Arnold et al., 2017; Scheuermann, 2017; Wahser & Sumfleth, 2008; Walpuski, 2006; Wollenschläger et al., 2012). Es fehlen also Studien, die sich mit denjenigen Schülerinnen und Schülern mit dem größten Förderbedarf auseinandersetzen.

3 Übergeordnete Fragestellungen

Kompetenzmodellierung und Aufgabenvalidierung

Gemäß Metzger (2013, 45) herrschen große Lücken zwischen dem, was in Standards beschrieben, mit Modellen modelliert und bei Assessments erhoben wird. Entsprechend war es ein Ziel dieser Dissertation, diese Lücke exemplarisch für den Problemtyp effektbasiertes Vergleichen zu schließen. Daraus ergeben sich die ersten beiden übergeordneten Fragestellungen:

1. *Inwiefern eignen sich für die Sekundarstufe I (Jahrgangsstufe 7–9) die a priori gesetzten Qualitätsstandards beim experimentellen Problemtyp effektbasiertes Vergleichen zur Beschreibung experimenteller Kompetenzen?*

2. *Wie valide lassen sich diese Qualitätsstandards mit Hilfe von hands-on Testaufgaben und dazugehörigen Kodiermanualen erfassen?*

Umsetzung in der Praxis

Naturwissenschaftliche Kompetenzmodelle, die primär als Messmodelle zu Testzwecken entwickelt wurden, haben oft kaum einen Nutzen für den Erwerb und den Aufbau von Kompetenzen im Unterricht (Bernholt, Parchmann & Commons, 2009, 219; Metzger, 2013, 46; Reusser, 2015, 94). Entsprechend war ein weiteres Ziel dieser Dissertation, basierend auf den Erkenntnissen der Aufgabenvalidierung, experimentelle Lerngelegenheiten für den naturwissenschaftlichen Unterricht zu gestalten, welche den Fokus auf die geforderten Kompetenzen (hier Qualitätsstandards) legen. Dies wurde im Rahmen dieser Dissertation exemplarisch am Problemtyp kategoriengeleitetes Beobachten dargestellt.

3. *Inwiefern können die Ergebnisse aus den ersten Pilot- resp. Validierungsstudien (bezogen auf die Progession entlang der Qualitätsstandards) beim Erstellen von Lerngelegenheiten bzw. Fördertools zum experimentellen Problemtyp kategoriengeleitetes Beobachten umgesetzt werden?*

Formative Beurteilung und Art der Lernunterstützung

Als weiteres Ziel dieser Dissertation sollte das Kompetenzmodell aus dem Projekt ExKoNawi auch als Diagnose- und Förderinstrument für die formative Beurteilung von Schülerinnen und Schülern im Bereich des praktisch-naturwissenschaftlichen Arbeitens genutzt werden:

4. *Inwiefern eignet sich das Kompetenzmodell aus dem Projekt ExKoNawi für den experimentellen Problemtyp effektbasiertes Vergleichen und die daraus abgeleiteten Aufgaben und adaptiven Fördermaßnahmen (hier kompetenzbezogenes Feedback), um Schülerinnen und Schüler der Sekundarstufe I bei der Entwicklung ihrer experimentellen Kompetenzen zu unterstützen?*

Die meisten naturwissenschaftsdidaktischen Interventionsstudien im Bereich experimenteller Kompetenzen berücksichtigen nicht die Schülerinnen und Schüler aus tiefen Anforderungsniveaus. Jedoch benötigen gerade diese Jugendlichen am meisten Unterstützung (siehe 2.9). Entsprechend war ein wichtiges Anliegen dieser Dissertation herauszufinden, welche Art einer (kompetenzbezogenen) Lernunterstützung (hier Fokus Feedbackform) bei Schülerinnen und Schülern aus leistungsschwachen Klassen am erfolgreichsten ist:

5. *Welche Feedbackform (feeding back, feeding forward oder beides) unterstützt Schülerinnen und Schülern aus leistungsschwachen Klassen der Jahrgangsstufe 7 am stärksten bei der Entwicklung experimenteller Kompetenzen?*

4 Publikationsbasierte Umsetzung der übergeordneten Fragestellungen

In diesem Kapitel werden die vier Publikationen, welche die verschiedenen Teilaspekte und Ziele der Arbeit darstellen und adressieren, zunächst kurz vorgestellt. Die ersten beiden wissenschaftlichen Veröffentlichungen umfassen die detaillierte Konzipierung und Validierung experimenteller hands-on Testaufgaben, die auf einer problemtypenbasierten Kompetenzmodellierung aufbauen. Der Fokus liegt hier exemplarisch auf Testaufgaben zum Problemtyp effektbasiertes Vergleichen. Das Erstellen von Testaufgaben sowie erste Ergebnisse aus den Validierungsstudien aus dem Projekt ExKoNawi stellen die Grundlage für eine praxisorientierte Gestaltung von Lernaufgaben zum Thema "kompetent Beobachten", welche in der dritten Publikation beschrieben wird. Die letzte Publikation berichtet von einer ersten Intervention, bei der die Wirkung von adaptivem, kompetenzbezogenem Feedback beim Lernen mit Aufgaben zum effektbasierten Vergleichen untersucht wurde.

4.1 Publikation 1: Beurteilung und Förderung experimenteller Kompetenzen mit Aufgaben zum effektbasierten Vergleichen.

Die erste Publikation befasst sich vor allem mit der ersten übergeordneten Fragestellung. Sie liefert zunächst einen Überblick über den Kompetenzbegriff in den Naturwissenschaften und legt den Fokus auf die fachspezifischen Kompetenzen, die beim praktisch-naturwissenschaftlichen Arbeiten (hier als Experimentieren deklariert) gemäß deutschen Bildungsstandards sowie dem Schweizer Lehrplan 21 am Ende der obligatorischen Schulzeit in den Jahrgangsstufen 7 bis 9 gefördert werden sollen. Hier werden zwei Ansätze der Modellierung experimenteller Kompetenzen (Teilprozessansatz vs. Problemtypenansatz) kurz verglichen und die Aufgabenkonstruktion mittels problemtypenbasierter Kompetenzmodellierung erläutert.

Um eine hohe Standardisierung bei der Validierung experimenteller hands-on Testaufgaben zu gewährleisten, wird hier u. a. noch einmal aufgezeigt, welche Überlegungen hinter dem Erstellen eines sogenannten Aufgabenstamms sowie der standardisierten Kodiermanuale stecken. Die Ergebnisse der Validierungsstudien zeigen, dass die a priori festgelegten Komplexitätsstufen der Kompetenzmodellierung (im weiteren Verlauf auch Qualitäts-standards genannt) eine Lernhierarchie bei diesem Aufgabentyp aufzeigen, welche unabhängig von der Jahrgangsstufe, vom Geschlecht (hier nicht aufgezeigt), vom Anforderungsniveau (Schulsystem), sowie vom Aufgabenkontext ist.

Eine überarbeitete online-Version dieser Artikels finden sie hier: https://onlinelibrary.wiley.com/doi/full/10.1002/ckon.201810322 (letzter Zugriff am 16.09.2019).

4.2 Publikation 2: Validating performance assessments: measures that may help to evaluate students' expertise in 'doing science'.

Diese Publikation befasst sich mit der zweiten übergeordneten Fragestellung und fokussiert die in der ersten Publikation bereits erwähnte (strukturelle) Validierung der experimentellen hands-on Testaufgaben, ergänzt diese jedoch um weitere Validitätsaspekte (Generalisierbarkeit und externe Validität).

Da die *Personen-Aufgaben Varianz* (siehe dazu auch *task-sampling variability* bei Gao, Shavelson & Baxter, 1994) bei experimentellen hands-on Testaufgaben (performance assessments) sehr hoch ist resp. die Leistung der Schülerinnen und Schüler bei solchen Tests sehr stark zwischen den eingesetzten Aufgaben variiert, wurde u. a. versucht, mittels hoher Standardisierungen (Aufgabenstamm, Kodiermanuale, Kodierertraining) und weiterer Maßnahmen (unterschiedliche Erhebungs- und Messverfahren, Erhöhung der Anzahl Testaufgaben/Person) die geforderte Güte (Validität) zu erreichen. Die Ergebnisse aus den unterschiedlichen Validierungsstudien werden quantitativ wie qualitativ beschrieben.

Hinweis: Eine weitere Publikation zur Validierung experimenteller hands-on Aufgaben (mit dem Fokus auf deren Generalisierbarkeit), welche nicht in diese Sammlung aufgenommen wurde, befindet sich im Tagungsband 38 der Gesellschaft für Didaktik der Chemie und Physik, anlässlich der Jahrestagung in Regensburg 2017:

Hild, P., Gut, C., Metzger, S. & Tardent, J. (2018). Zur Generalisierbarkeit bei Experimentiertests. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht – normative und empirische Dimensionen* (S. 348–351). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Regensburg 2017. http://www.gdcp.de/index.php/tagungsbaende/tagungsband-uebersicht/167-tagungsbaende/2018/10912-concatenate-d37-a37-10912 (letzter Zugriff am 18.05.2019)

4.3 Publikation 3: Beobachten lernen. Aufgaben zur Förderung der Beobachtungskompetenz.

Die dritte Publikation widmet sich der dritten übergeordneten Fragestellung. Bei der Modellentwicklung müssen auch unterrichtspraktische Notwendigkeiten, wie die Anschlussfähigkeit und Nützlichkeit des Kompetenzmodells für Unterrichtsgestaltung und Individualdiagnostik, Rechnung getragen werden (z. B. Bernholt, Parchmann & Commons, 2009). Diesem Desiderat wurde im Projekt ExKoNawi zum Teil nachgegangen, indem das entstandene Kompetenzmodell (für das praktisch-naturwissenschaftliche Arbeiten) als Planungs- (wie auch Diagnose-)instrument für die Schulpraxis konzipiert wurde und bereits jetzt in der Lehrpersonenausbildung bei der Vermittlung fachdidaktischer Inhalte seine Verwendung findet (Ausbildung Sekundarstufe I, Pädagogische Hochschule Zürich).

Im Rahmen dieser Dissertation entstanden mehrere Fachartikel und Unterrichtseinheiten, welche sich ganz bewusst auf die Kompetenzmodellierung aus dem Projekt ExKoNawi stützen. Zum Beispiel widmet sich Publikation 3 der Frage, wie Schülerinnen und Schüler kompetent beobachten lernen (Problemtyp kategoriengeleitetes Beobachten) und liefert hierzu konkrete Tools für den Unterricht (Abb. 7.2). Die beiden vorgestellten Lernaufgaben mit chemischen Kontexten (pH-Abhängigkeit einer Tintensorte sowie unterschiedliches Verhalten von Öl-in-Wasser- vs. Wasser-in-Öl-Emulsionen) wurden demnach, basierend auf den Qualitätsstandards zum Problemtyp kategoriengeleitetes Beobachten erstellt. Zudem wird den Lehrpersonen wie auch den Schülerinnen und Schülern eine Methodenkarte zur Verfügung gestellt.

Weitere Lernaufgaben mit Fokus Chemie basierend auf den Ergebnissen der ersten Pilotierungen aus dem Projekt ExKoNawi und der Dissertation wurden im Raabits Verlag veröffentlicht:

Hild, P. & Kallinna, K. (2018). Heavy Metal & Co. *RAAbits Naturwissenschaften, 27*. Dr. Josef Raabe Verlags-GmbH.

Hild, P. & Kölbach E. (2017). Der Schminkkoffer der Alten Ägypter, *RAAbits Naturwissenschaften, 22.* Dr. Josef Raabe Verlags-GmbH.

Kölbach, E. & Hild, P. (2015). Ist Tee gleich Tee? – Unterscheidung von Tee und Aufgussgetränken mit Schülerexperimenten. *RAAbits Chemie, 30*. Dr. Josef Raabe Verlags-GmbH.

4.4 Publikation 4: Adaptives kompetenzbezogenes Feedback beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten. Eine empirische Untersuchung zur Wirksamkeit unterschiedlicher Feedbackformen.

Die letzte Publikation widmet sich den beiden letzten übergeordneten Fragestellungen (4 und 5). In einem ersten Schritt werden unterschiedliche Scaffolding-Maßnahmen beim praktisch-naturwissenschaftlichen Arbeiten beschrieben und Forschungsergebnisse bezogen auf das selbstständige Lernen beim praktisch-naturwissenschaftlichen Arbeiten verglichen. In der Intervention wird die Wirksamkeit unterschiedlicher Feedbackformen ((rückmeldendes) feeding back vs. (hinweisgebendes) feeding forward oder beides gleichzeitig, vgl. Hattie & Timperley, 2007) auf den Zuwachs der *experimentellen Kompetenz*[2] von Schülerinnen und Schüler aus leistungsschwachen Klassen der Jahrgangsstufe 7 untersucht.
Die Beurteilung der Kompetenz und das daran gekoppelte adaptive Feedback beziehen sich dabei auf das im Vorfeld validierte problemtypenbasierte Kompetenzstufenmodell zum effektbasierten Vergleichen (Publikationen 1 und 2).

[2] In dieser Publikation wird *experimentelle Kompetenz* im Singular verwendet, da es sich hier um eine Messvariable handelt.

5 Beurteilung und Förderung experimenteller Kompetenzen mit Aufgaben zum effektbasierten Vergleichen

Hild, P., Metzger, S. & Parchmann, I. (2018). Beurteilung und Förderung experimenteller Kompetenzen mit Aufgaben zum effektbasierten Vergleichen. *Chemkon, 25*(3), 90–97.

5.1 Zusammenfassung

Dieser Artikel liefert konkrete Beispiele zur Beurteilung und Förderung experimenteller Kompetenzen im Bereich Erkenntnisgewinnung. Es werden zunächst Grundlagen für die Entwicklung kompetenzorientierter hands-on Testaufgaben am Beispiel des effektbasierten Vergleichens beschrieben. Bei diesen Aufgaben müssen Schülerinnen und Schüler unterschiedliche Objekte anhand einer gegebenen Eigenschaft experimentell vergleichen. Unterschiedliche zu erreichende Kompetenzen aus dem Bereich der Erkenntnisgewinnung wurden a priori, anhand eines Strukturmodells, postuliert. Mithilfe validierter Testaufgaben und eines standardisierten Manuals konnte bei einer Stichprobe von 418 12- bis 15-jährigen Schülerinnen und Schülern (jeweils 2 Aufgaben mit gleichem Aufgabenstamm pro Person) eine Hierarchie entlang dieser Kompetenzen nachgewiesen werden. Die Übereinstimmungen zwischen den Prüferinnen und Prüfern liegen alle in einem akzeptablen bis sehr guten Bereich. Die Ergebnisse sollen die Grundlage für die Konstruktion von Lernumgebungen liefern, die es den Lehrpersonen erlauben, Kompetenzen im Bereich der Erkenntnisgewinnung zu beurteilen und individuell zu fördern.

5.2 Einleitung

In diesem Artikel werden Aufgaben zum effektbasierten Vergleichen vorgestellt und es wird aufgezeigt, wie Lehrpersonen Feedback hinsichtlich der zu erreichenden Kompetenzen geben können.

Aufgaben spielen im Unterricht eine zentrale Rolle und sind wichtige Mediatoren zwischen den Anforderungen von Lehrplänen und dem einsetzenden schulischen Lernerfolg (Ellett, 1986). Nachdem durch erste Ergebnisse aus den TIMS und PISA Studien deutlich wurde, dass Schülerinnen und Schüler Schwierigkeiten mit der Bearbeitung von Aufgaben haben, bei denen das konzeptuelle Verständnis, das selbstständige Anwenden und Übertragen des Gelernten auf neue Kontexte sowie das flexible Umstrukturieren von Problemkonstellationen erforderlich sind, kam der Wunsch nach einer neuen Aufgabenkultur auf (Baumert, Bos & Lehmann, 2000). Diese beinhaltet eine Steigerung der fachdidaktischen Qualität allgemein sowie die Kompetenzorientierung der Aufgaben (Heitzmann, 2012).

Im deutschsprachigen Raum sind seit einigen Jahren die Beschreibungen von Kompetenzen, die Standards, an Auflagen gebunden, welche sich geschichtlich und kulturell sowie auch durch unterschiedliche induktive oder deduktive Vorgehensweisen bei deren Erstellung voneinander unterscheiden. In der Schweiz wurden die Standards für das Fach Natur und Technik (Biologie, Chemie und Physik in einem Fach) deduktiv, basierend auf nationalen “Papier und Bleistift”– sowie Experimentiertests ausformuliert (Schweizer Konferent der kantonalen Erziehungsdirektoren, 2011). Für den Bereich Fragen und Untersuchen (Erkenntnisgewinnung), wurden unter anderem Schülerbeiträge, die während und nach dem Lösen von problemorientierten hands-on Testaufgaben erstellt wurden, analysiert. Progressionen entlang der Standards konnten ausformuliert, jedoch statistisch nicht nachgewiesen werden (Gut, 2012): Die Testaufgaben scheinen unterschiedliche Problemlösestrategien und unterschiedliche Expertisen von den Schülerinnen und Schülern zu verlangen. Im Projekt ExKoNawi der PH Zürich wurde ein Testinstrument entwickelt und validiert, das unterschiedlichen experimentellen Problemtypen unterschiedliche Kompetenzstrukturen und -progressionen zuordnet (Gut et al., 2014).

Einer dieser experimentellen Problemtypen, das effektbasierte Vergleichen, wird im dritten Teil genauer vorgestellt und untersucht.

5.3 Kompetenzorientierung

5.3.1 Der Kompetenzbegriff in den Naturwissenschaften

Die Kompetenzdiskussion im deutschsprachigen Raum wurde im Wesentlichen von der ganzheitlichen Definition von Weinert geprägt: Kompetenzen sind „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbunden motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (Weinert, 1999). Auch für die Schweizer Grundkompetenzen und den Lehrplan 21 diente sie als Basis (vgl. Konsortium Harmos Naturwissenschaften+, 2010; D-EDK, 2015). Dennoch werden in den naturwissenschaftlichen Kompetenzformulierungen explizit fast ausschließlich kognitive Fähigkeiten und Fertigkeiten beschrieben. Motivationale, volitionale und soziale Fähigkeiten und Bereitschaften erscheinen dagegen – wenn überhaupt nur implizit. Dies entspricht eher der auf die fachspezifischen Kompetenzen fokussierten Kompetenzdefinition von Meyer (2009): „Eine Kompetenz bezeichnet die Fähigkeit, durch Erfahrungen und Lernen erworbenes Wissen und Können in immer wieder neuen Handlungssituationen selbstständig, verantwortungsbewusst und situationsangemessen anzuwenden."

In Tabelle 5.1 sind verschiedene Facetten der Kompetenzdefinitionen zusammengestellt und mit Beispielen aus den deutschen Bildungsstandards (Chemie) und dem Schweizer Lehrplan21 (Natur und Technik) illustriert. Es wird deutlich, dass kognitive Fähigkeiten und Fertigkeiten kaum getrennt werden können: Zu handeln, also etwas zu tun, ist ohne Wissen in der Regel nicht möglich. Außerdem zeigt ein Blick in die kompetenzorientierten Ausformulierungen bei beiden Ländern, dass eine Unterscheidung der ersten drei Teilbereiche der Scientific Literacy nach Hodson (2009) kaum möglich zu sein scheint. Zumindest enthalten die dort aufgeführten Standards jeweils inhaltliche (knowing science), prozedurale (doing science) und epistemische Wissenselemente (knowing about science). Der vierte Teilbereich (engaging in sociopolitical action) dagegen kann separat dargestellt und den überfachlichen Kompetenzen zugeordnet werden. Allerdings bleibt festzuhalten, dass alle diese Kompetenzfacetten nicht trennscharf sind und sich jeweils gegenseitig beeinflussen.

Tab. 5.1 Kompetenzbegriff in den Naturwissenschaften.

Kompetenzbegriff nach Weinert (1999)	kognitive Fähigkeiten und Fertigkeiten	motivationale, volitionale und soziale Bereitschaften & Fähigkeiten
Kompetenzen (allgemein)	fachspezifische Kompetenzen	…auf das Fach bezogene überfachliche Kompetenzen
***scientific literacy* nach Hodson** (2009)	knowing science doing science knowing about science	engaging in sociopolitical action
Bildungsstandards für den Mittleren Schulabschluss (KMK, 2004)	Fachwissen Erkenntnisgewinnung Bewerten	Kommunikation
Grundkompetenzen Naturwissen-schaften+ (EDK, 2011)	Informationen erschließen Fragen und Untersuchen Ordnen, Strukturieren und Modellieren Einschätzen und Beurteilen Entwickeln und Umsetzen	Interesse und Neugierde entwickeln Eigenständig Arbeiten Mitteilen und Austauschen
Beispiele aus dem Schweizer Lehrplan21 (D-EDK, 2009)	Schülerinnen & Schüler … können einfache Gemische mit ausgewählten Methoden nach Anleitung trennen und das Vorgehen fachlich korrekt beschreiben (NT.2.2.b).	Schülerinnen & Schüler ...können verschiedene Formen der Gruppenarbeit anwenden. … können unterschiedliche Sachverhalte sprachlich ausdrücken und sich dabei anderen verständlich machen.
	SchülerInnen können selbstständig in Medien nach Informationen zum Recycling von Stoffen suchen und das eigene Recyclingverhalten reflektieren (NT.3.3.d1).	
Beispiele aus den deutschen Bildungs-standards im Fach Chemie (KMK, 2004)	Schülerinnen & Schüler ...planen geeignete Untersuchungen zur Über-prüfung von Vermutungen und Hypothesen (E2).	Schülerinnen & Schüler … diskutieren und bewerten gesellschaftsrelevante Aussagen aus unterschiedlichen Perspektiven (B5).
	SchülerInnen planen, strukturieren, reflektieren und präsentieren ihre Arbeit als Team (K10).	

5.3.2 Fachspezifische Kompetenzen beim Experimentieren

Die Naturwissenschaften unterscheiden sich von anderen Disziplinen unter anderem durch die Art und Weise der Erkenntnisgewinnung (im Sinne von Nature of Science). Dies zeigt sich letztlich auch in den Ausformulierungen der Kompetenzen. In den letzten Jahren wurden gerade für den Kompetenzbereich der naturwissenschaftlichen Erkenntnisgewinnung unterschiedliche Struktur- und Progressionsmodelle (Schecker & Parchmann, 2006) postuliert und zum Teil validiert, die sich vor allem mit der Rolle des Experimentierens befassen (Hammann, Phan & Bayrhuber, 2007; Mayer et al., 2008; Neumann et al., 2007).

Experimentieren wird häufig als Zusammenspiel unterscheidbarer Teilprozesse wie z. B. Hypothesen aufstellen, Experimente planen, Daten auswerten, ... verstanden (Emden, 2011). Diese Teilprozesse können je nach Experimentiersituation unterschiedlich stark gewichtet und gefördert werden. Andererseits kann Experimentieren auch als integraler Problemlöseprozess angesehen werden, wobei Lernende je nach Problemstellung unterschiedliche Kompetenzen benötigen und neues Wissen im Prozess der Problembearbeitung generieren (Gott & Duggan, 1996; Reusser, 2005). So können reine Beobachtungen mit dem letzten Ansatz schon als eigenständiges Experiment verstanden werden. Die beiden Ansätze unterscheiden sich in ihren Anforderungen an die Schülerinnen und Schüler (Rumann et al., 2010). Bevor individuelle Lernprozesse und/oder Lernprodukte einer Aufgabe im Bereich Erkenntnisgewinnung untersucht werden können, muss deshalb zuerst geklärt werden, welcher Ansatz gewählt wird.

5.3.3 Der Problemlöseansatz

Um die Vielfalt an experimentellen Aktivitäten im Unterricht umfassend zu beschreiben, bietet sich für die Aufgabenkonstruktion sowie für die Kompetenzbeurteilung der Problemlöseansatz aus dem Projekt ExKoNawi an (Gut et al., 2014): Schülerinnen und Schüler haben dann einen Zuwachs experimenteller Kompetenzen, wenn sie (hier in den Lernprodukten) komplexere Probleme qualitativ besser, eigenständiger und in mehr fachlichen Kontexten lösen oder ihre Leistungen in wiederholten Situationen stabiler werden. Dabei wird zwischen verschiedenen experimentellen Problemtypen wie beispielsweise kategoriengeleitetes Beobachten, skalenbasiertes Messen, fragengeleitetes Untersuchen und effektbasiertes Vergleichen unterschieden. Tabelle 2.1 zeigt die zu diesen Problemtypen gehörenden fachspezifischen Kompetenzbe-schreibungen, die in naturwissenschaftlichen Lernumgebungen gefördert und überprüft werden können.

5.4 Aufgabenkonstruktion

5.4.1 Aufgaben zum effektbasierten Vergleichen

Aufgaben zum experimentellen Problemtyp effektbasiertes Vergleichen werden definiert als Aufgaben, bei denen Objekte anhand einer gegebenen Eigenschaft verglichen werden sollen und alle anderen Variablen konstant gehalten werden (Solano-Flores & Shavelson, 1997). Neben den Validierungsstudien innerhalb des Projekts ExKoNawi und der large-scale Leistungsmessungen von Shavelson et al. (1998) wurde dieser experimentelle Problemtyp, bis dato, kaum untersucht (vgl. Hungerfold & Miles, 1969; Meyer & Carlisle, 1996; Tomera, 1974) – obwohl das effektbasierte Vergleichen in vielen Problemlösesituationen im Unterricht sowie in internationalen Vergleichsstudien explizit gefordert wird (Harmon et al., 1997; Stecher, 1996). Schülerinnen und Schüler sollen bei diesem Problemtyp Objekteigenschaften, wie zum Beispiel den Säuregehalt, die Löslichkeit, die wässrigen Eigenschaften, die Energieeffizienz, den osmotischen Effekt oder auch die magnetische Kraft unterschiedlicher Objekte vergleichen. Dabei wird zwischen qualitativen Vergleichen (Reihenfolge festlegen vom stärksten zum schwächsten) und quantitativen Vergleichen (welche Objekte sind sich am ähnlichsten) unterschieden.

Tab. 5.2 Unterschiedliche Kontexte beim effektbasierten Vergleichen.

Objekte	**Schülerinnen und Schüler vergleichen…**
Kartoffeln	*…den osmotischen Effekt unterschiedlicher Pulver (Zucker, Mehl, Salz) auf Kartoffeln.*
Magnete	*…die Stärke unterschiedlicher Magnete.*
Milchprodukte	*…die wässrigen Eigenschaften von Emulsionen (Magerquark, Streichkäse).*
Pulver	*…die Löslichkeit unterschiedlicher Pulver (Einfach- & Zweifachzucker, Kochsalz rein) in Wasser.*
Säfte	*…die Menge Wasser in unterschiedlichen Früchten (Zucchini, Apfel, Gurke).*
Solarzellen	*…die Effizienz unterschiedlicher Solarzellen.*
Nüsse (Video)	*...den Fettgehalt unterschiedlicher Steinfrüchte und Nüsse (Erdnüsse, Walnüsse, Mandeln).*
Tee (Video)	*...den Säuregehalt unterschiedlicher Infusionen (Rooibos, Früchte, Schwarztee).*

Im Projekt ExKoNawi wurde dieser Problemtyp in sechs unterschiedlichen Kontexten (siehe Tabelle 5.2) untersucht. Zwei weitere Aufgaben (Nüsse und Tee) wurden für weitere Validierungsstudien erstellt.

5.4.2 Aufgabenstamm

Eine genügend hohe Reliabilität bei Leistungsmessungen in Testaufgaben mit unterschiedlichen Kontexten kann durch standardisierte Aufgabenstellungen, so genannte Aufgabenstämme, gewährleistet werden (Solano-Flores et al., 1999). Jedoch belegen keine Studien, dass dadurch auch die hohe Varianz zwischen den einzelnen Testaufgaben reduziert sowie deren Auswechselbarkeit gewährleistet werden können (Ruiz-Primo, Baxter & Shavelson, 1993) .

Tabelle 5.3 zeigt, wie Aufgaben zum effektbasierten Vergleichen im Projekt ExKoNawi aufgebaut wurden: Die Schülerinnen und Schüler müssen in jeweils 18 Minuten Objekte anhand einer gegebenen Eigenschaft vergleichen und ihre Beobachtungen, Reflexionen und Resultate protokollieren. Sie werden dabei durch den Prozess geleitet, indem sie nacheinander vier Teilaufgaben lösen. Eine Beispielaufgabe befindet sich am Ende des Kapitels 3 (Tabelle 5.5). Der Einsatz von Vignetten (grau) hat sich hierbei bewährt, da sich gezeigt hat, dass dadurch Lernprozesse aktiviert wurden und gleichzeitig genauer protokolliert wurde. Alle Aufgaben wurden mit dem gleichen Aufgabenstamm entwickelt.

Tab. 5.3 Aufgabenstamm für effektbasiertes Vergleichen.

Titel	*[ein Begriff]*
Material	*[Bild mit Materialbeschriftung]*
Problem	*[**w**eiblicher **N**ame]* & *[**m**ännlicher **N**ame]* wollen herausfinden, welches *[**O**bjekt]* die beste *[**E**igenschaft]* hat. Sie haben jedoch keine Ahnung, wie sie das machen sollen. Bearbeite dazu auf den folgenden Seiten **4 Aufgaben**.
A1	Vergleiche die *[O]* ***A*** und ***B***. Finde durch Experimentieren heraus, welches *[O]* die beste *[E]* hat. Was hast du herausgefunden? Kreuze an. ☐ **A** ist *[E]* als **B** ☐ **B** ist *[E]* als **A**
	[w. N.] & *[m. N.]* haben ein anderes Resultat erhalten. Sie wollen wissen, wie du auf dein Resultat kommst. Beschreibe und skizziere, welche Überlegungen, Experimente und Beobachtungen du gemacht hast. Erkläre es so, dass *[w. N.]* & *[m. N.]* deinen Versuch selber durchführen können.
A2	Untersuche nun auch noch ***C***. Erstelle eine Reihenfolge der *[O]* **A**, **B** und **C**. Beginne mit dem *[O]*, das die besten *[E]* hat.
	[w. N.] & *[m. N.]* haben eine andere Reihenfolge erhalten. Sie verstehen nicht, wie du auf deine Reihenfolge kommst. Beschreibe und skizziere, welche Experimente und Beobachtungen du gemacht hast. Erkläre es so, dass *[w. N.]* & *[m. N.]* deinen Vergleich selber durchführen können.
A3	Worauf hast du geachtet, dass deine Vergleiche fair sind?
A4	Untersuche nun noch, welche zwei *[O]* sich am ähnlichsten sind, wenn man die *[E]* untersucht. Wenn nötig, führe weitere Vergleiche durch, um das Problem zu lösen. Was hast du herausgefunden? Kreuze an. ☐ **A** und **B** sind sich am ähnlichsten ☐ **A** und **C** sind sich am ähnlichsten ☐ **B** und **C** sind sich am ähnlichsten
	[w. N.] & *[m. N.]* haben eine andere Lösung erhalten. Sie verstehen nicht, wie du auf deine Lösung kommst. Beschreibe und skizziere, welche Vergleiche und Beobachtungen du gemacht hast. Erkläre es so, dass *[w. N.]* & *[m. N.]* deinen Vergleich selber durchführen können.

5.5 Kompetenzbeurteilung

5.5.1 Manual

Um die facettenreichen Aktivitäten des Experimentierens zu bewerten, können Lösungsprozesse wie auch Lösungsprodukte analysiert werden. Im Projekt ExKoNawi wurde nur das Produkt, die Schülerprotokolle, anhand eines Manuals bewertet. Basierend auf den Ergebnissen und Desiderata früherer Studien (vgl. Solano-Flores & Shavelson, 1997; Shavelson, Solano-Flores & Ruiz-Primo, 1998; Hungerfold & Miles, 1969; Meyer & Carlisle, 1996; Tomera, 1974), wurden 4 Kompetenzen (bestehend aus jeweils eins bis drei Kriterien) ausformuliert, die von den Schülerinnen und Schülern bei diesem Problemtyp erreicht werden sollen (kursiv konkrete Beispiele aus dem Manual):

- *K1: Einen theoretisch funktionierenden Vergleich durchführen. Der Vergleich ist adäquat und wird in mindestens qualitativer Form beschrieben. Er muss nicht fair sein.*
- *K2: Eine qualitative Reihenfolge von drei Objekten aufstellen. Es wird offensichtlich mit allen Objekten experimentiert. Eine korrekte Reihenfolge wurde erstellt.*
- *K3: Bedingungen für einen fairen Vergleich beschreiben. Es werden explizit mehrere Variablen erwähnt, die bei diesem Vergleich kontrolliert wurden.*
- *K4: Objekte mithilfe quantitativer Aussagen vergleichen. Die Vergleiche lassen quantitative Aussagen über die Ausprägungen oder über Differenzen der Ausprägungen der Objekteigenschaft zu.*

Mit den verschiedenen Teilaufgaben im Aufgabenstamm, lassen sich diese Kompetenzen überprüfen. Eine Kompetenz wurde als “erreicht” eingestuft, wenn > 50% der Kriterien im Protokoll erfüllt waren (Gut et al., 2014). Anschließend konnte die a priori festgelegte Lernhierarchie entlang dieser vier Kompetenzen (K) untersucht werden (Abbildung 5.1).

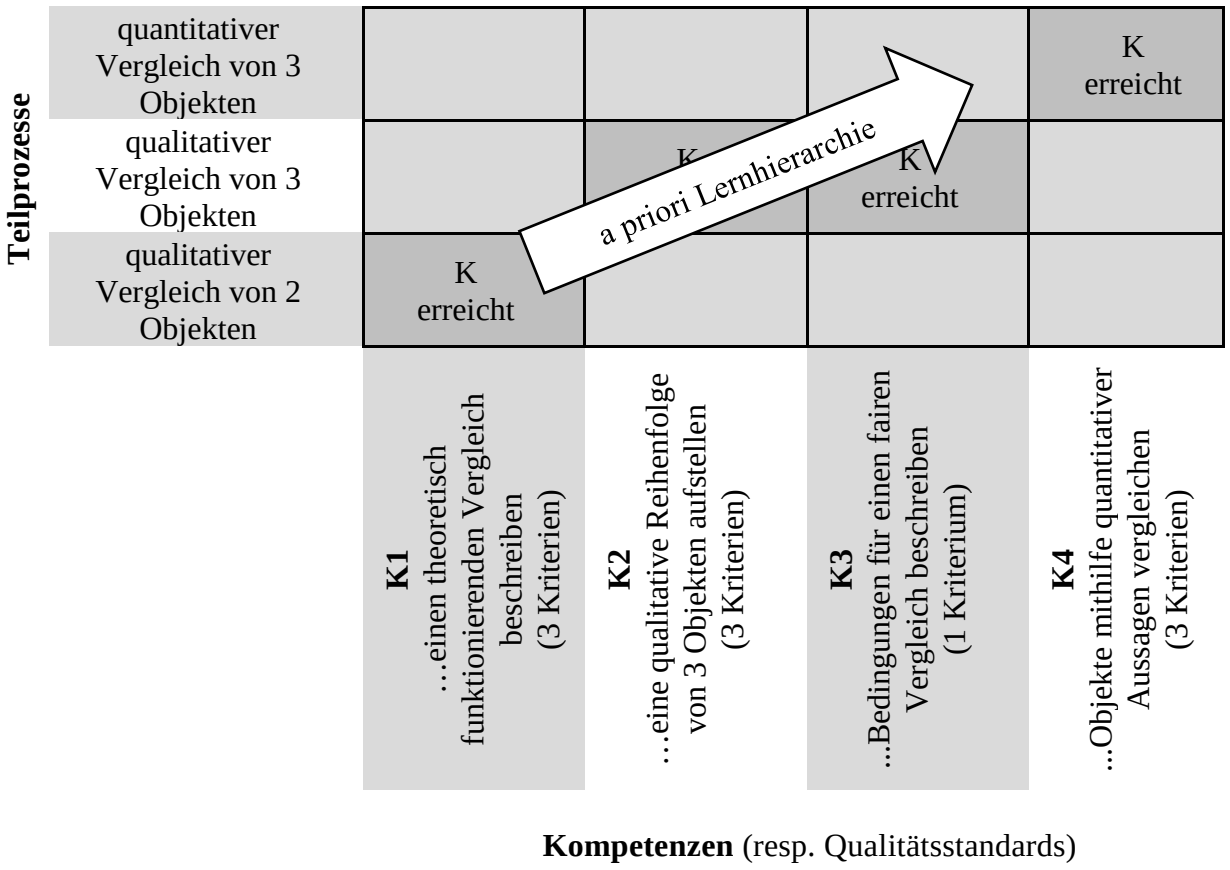


Abb. 5.1 Vier Kompetenzen beim effektbasierten Vergleichen.

5.5.2 Hierarchie entlang der Kompetenzen

Um die postulierte Lernhierarchie entlang der 4 Kompetenzen zu überprüfen, lösten und protokollierten 418 Schülerinnen (49%) und Schülern der Sekundarstufe I (12–15 Jahre alt) im Kanton Zürich, zu drei unterschiedlichen Erhebungszeitpunkten, jeweils zwei Aufgaben (Tabelle 5.4, BC bezeichnet das tiefste, A das mittlere Niveau im Schweizer Schulsystem der 12- bis 15-Jährigen).

Tab. 5.4 Informationen zur Stichprobe.

Messzeitpunkt	**n**	**7te**		**8te**	**9te**	
		BC	**A**	**BC**	**BC**	**9A**
I	152	*26*	*44*	*17*	*24*	*41*
II	190	*54*	*63*		*23*	*50*
III	76				*40*	*36*

Anmerkung: n = Stichprobengröße

Alle Protokolle wurden von zwei Personen ausgewertet. Die Übereinstimmungen zwischen den Personen bei der Auswertung der Schülerprotokolle liegen in einem akzeptablen bis sehr guten Bereich ($.56 \leq \kappa \leq .97$; $.79\ p_0 \leq .98$). Obwohl die Aufgaben unterschiedlich gut zu den drei verschiedenen Messzeitpunkten (mit unterschiedlichen Personen) gelöst wurden, konnte die postulierte Hierarchie (K1 wird von mehr Jugendlichen erreicht als K2, K2 von mehr als K3 und K3 von mehr als K4) in allen Erhebungen bei allen Aufgaben nachgewiesen werden. Da die beiden Aufgaben Nüsse und Tee nur von 16 Schülerinnen und Schülern durchgeführt wurden, fehlen uns Daten für diese beiden weiteren Kontexte. Die relativen Häufigkeiten für das Erreichen der Kompetenzen K1 bis K4 unterscheiden sich (Wilcoxon Test) bis auf eine Ausnahme jeweils (bei 3 Messzeitpunkten und 3 Aufgaben) signifikant (Abbildungen 5.2, 5.3 und 5.4). Die vierte Kompetenz wurde fast nie erreicht und aus diesem Grund nicht in den Abb. 5.2 bis 5.4 dargestellt.

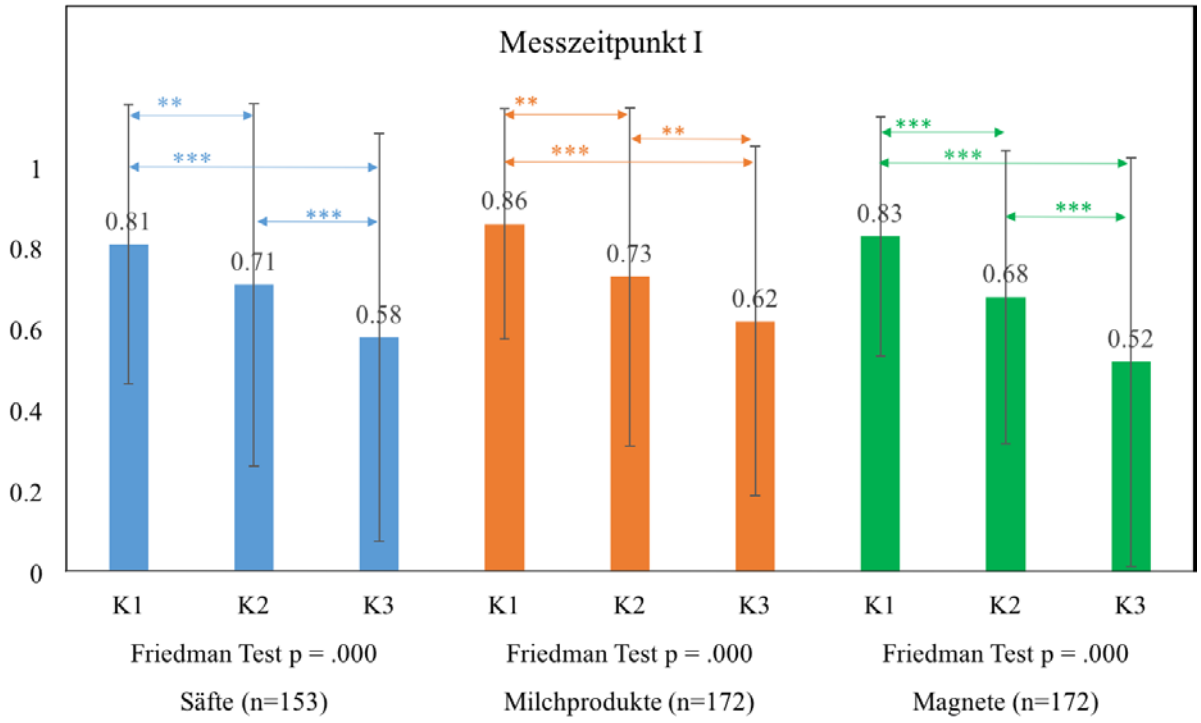


Abb. 5.2 Lernhierarchie beim Messzeitpunkt I (n=152).

Ähnlich wie in den Studien von Gut (2012) oder auch Gao, Shavelson und Baxter (1994), zeigte die Analyse der Daten jedoch auch, dass die unterschiedlichen Testaufgaben immer noch zu wenig korrelieren (Spearman $\rho < .6$). Höhere Korrelationen sollten durch weitere Standardisierungen der Testhefte erreicht werden können.

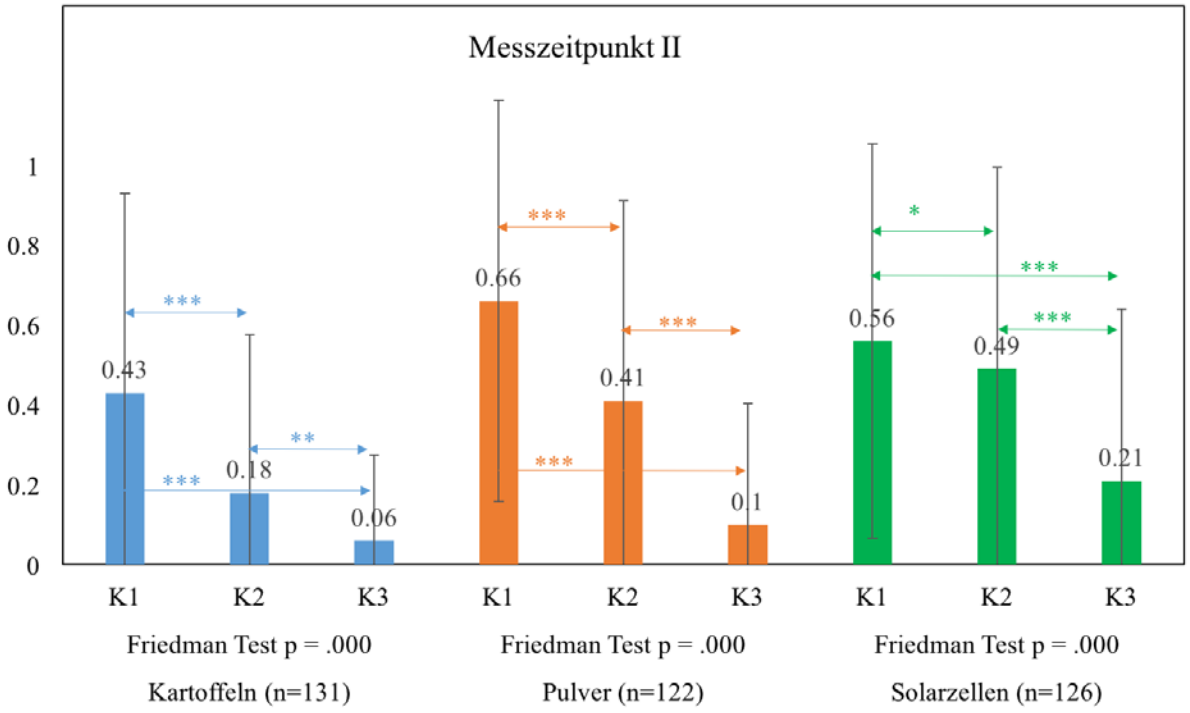


Abb. 5.3 Lernhierarchie beim Messzeitpunkt II (n=190).

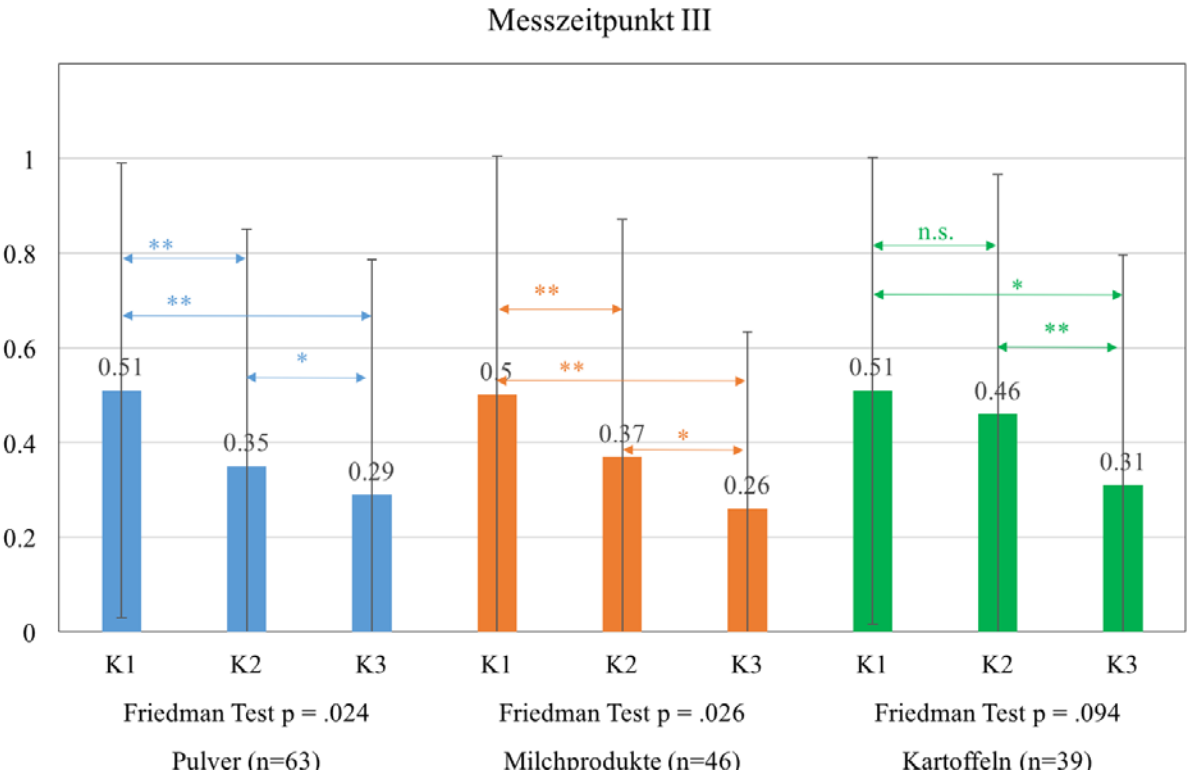


Abb. 5.4 Lernhierarchie beim Messzeitpunkt III (n=76).

5.6 Kompetenzförderung

Durch die Beurteilung der gezeigten Leistungen soll eine Lernunterstützung und Begleitung der Lernprozesse und damit die Gesamtkonzeption tragfähiger kompetenzorientierter Lernumgebungen ermöglicht werden (Oelkers, 2012).

5.6.1 Feeding back und feeding forward

Generell kann zwischen rückmeldenden und hinweisgebenden Angeboten unterschieden werden: feeding back und feeding forward (Hattie & Timperley, 2007). Feeding back erfolgt lerndiagnostisch in Form formativer (meist korrektiver) Angaben (William, 2011). Feeding back ist adaptiv, wenn die Aufgaben, wie hier, individuelle Kompetenzen und Kompetenzzuwächse aufzeigen können. Feeding forward Angebote sind häufig Lösungshinweise und haben einen lernfördernden Charakter. Sie können Lernhilfen oder Lösungsbeispiele (worked-out examples) zur individuellen Unterstützung des Bearbeitungsprozesses enthalten (z. B. Kölbach, Maier-Richter & Sumfleth, 2014). Falls feeding forward Angebote adaptiv sind, zählen sie zu den Scaffolding-Maßnahmen (vgl. Krammer, 2009; Luthiger et al., 2014, van de Pol et al., 2010). Sind solche Angebote gestuft und werden zum Beispiel von der Lehrperson erst bei Bedarf ausgeteilt (vgl. Stäudel et al., 2010), so kann diese Adaptivität im Lernfortschritt gewährleistet werden.

Ein mögliches feeding back zu Aufgaben zum effektbasierten Vergleichen könnten Hinweise zum Erreichen der vier Kompetenzen sein. Feeding forward könnte in Form von Lernhilfen (z. B. Hinweise, was einen fairen Vergleich ausmacht) oder Lösungsbeispielen (z. B. zum korrekten Arbeiten mit dem Material) angeboten werden. Standardisierte Feeding back- und feeding forward-Angebote zur Aufgabe in Tabelle 5.5 befinden sich in Tabelle 5.6 am Ende des Kapitels 5.

Der gleichzeitige Einsatz von feeding back und feeding forward wurde in mehreren Studien untersucht (vgl. Bloom, 1984; Gick & Holyoak, 1980; Lysakowski & Walberg, 1982). In einigen Studien konnte gezeigt werden, dass feeding forward einen größeren Einfluss auf den Lernerfolg hat als feeding back (Perfetto, Bransford & Franks, 1983; Vollmeyer & Rheinberg, 2005; Harks et al., 2014a; Scheuermann, 2017). Des Weiteren konnte gezeigt werden, dass Schülerinnen und Schüler besser lernen, wenn feeding back zur erreichten Kompetenz durch die Lehrperson und nicht durch eine Selbsteinschätzung erfolgt (Wollenschläger et al., 2012).

5.6.2 Ausblick

Aktuell fehlen naturwissenschaftsdidaktische Studien, die unterschiedliche Feedbackformen (feeding back, feeding forward & feeding up) vergleichen. Deshalb werden zurzeit in einem Folgeprojekt mögliche Kompetenzzuwächse dank feeding back und/oder feeding forward Angeboten bei Aufgaben zum effektbasierten Vergleichen in einer Interventionsstudie mit 12 Sekundarschulklassen (der siebten Klasse im Kanton Zürich) untersucht. Das Feedback soll hier auf den gemessenen Lernhierarchien (siehe Abb. 5.2 bis 5.4) aufbauen und der Lernstand anhand der validierten Testaufgaben kontinuierlich überprüft werden. In der Studie werden, nach einem Prätest, vier Gruppen gebildet: In der ersten Gruppe erhalten Schülerinnen und Schüler Aufgaben mit adaptierten feeding back und in der zweiten, Aufgaben mit adaptierten feeding forward Angeboten. Die Personen aus der dritten Experimentalgruppe erhalten beides. Mitglieder aus diesen drei Gruppen sollten im Vergleich (Post-Test) zu Personen aus der Kontrollgruppe, die während der Intervention kein adaptives Feedback erhalten, einen größeren Lernzuwachs haben. Die Ergebnisse sollen die Grundlage für die Konstruktion von Lernumgebungen liefern, die es den Lehrpersonen erlauben, Kompetenzen im Bereich Erkenntnisgewinnung zu beurteilen und individuell zu fördern.

Tab. 5.5 Beispielaufgabe Nüsse.

Titel	**Nüsse**
Material	6 Filterpapiere Erdnüsse Walnüsse Mandeln Reibe  Mörser mit Pistill Haushaltspapier Spatellöffel Bleistift
Problem	Xenia und Xaver wollen herausfinden, welche Nuss am meisten Fett enthält. Sie haben jedoch keine Ahnung, wie sie das machen sollen. Bearbeite dazu auf den folgenden Seiten **4 Aufgaben**.
A1	Vergleiche die **Erdnüsse** mit den **Walnüssen**. Finde durch Experimentieren heraus, welche Nuss den größten Fettfleck hinterlässt. Was hast du herausgefunden? Kreuze an. ☐ Erdnüsse hinterlassen einen größeren Fettfleck als Walnüsse. ☐ Walnüsse hinterlassen einen größeren Fettfleck als Erdnüsse.
	Xenia und Xaver haben ein anderes Resultat erhalten. Sie wollen wissen, wie du auf dein Resultat kommst. Beschreibe und skizziere, welche Überlegungen, Experimente und Beobachtungen du gemacht hast. Erkläre es so, dass Xenia und Xaver deinen Vergleich selber durchführen können.
A2	Untersuche nun auch noch die **Mandeln**. Erstelle eine Reihenfolge der Nüsse. Beginne mit der Nuss, die den größten Fettfleck hinterlässt
	Xenia und Xaver haben eine andere Reihenfolge erhalten. Sie verstehen nicht, wie du auf deine Reihenfolge kommst. Beschreibe und skizziere, welche Experimente und Beobachtungen du gemacht hast. Erkläre es so, dass Xenia und Xaver deinen Vergleich selber durchführen können.
A3	Worauf hast du geachtet, dass deine Vergleiche fair sind?
A4	Untersuche nun noch, welche zwei Nüsse sich am ähnlichsten sind, wenn man den Fettfleck untersucht. Wenn nötig, führe weitere Vergleiche durch, um das Problem zu lösen. Was hast du herausgefunden? Kreuze an. ☐ **Erdnüsse** und **Walnüsse** sind sich am ähnlichsten ☐ **Erdnüsse** und **Mandeln** sind sich am ähnlichsten ☐ **Walnüsse** und **Mandeln** sind sich am ähnlichsten
	Xenia & Xaver haben eine andere Lösung erhalten. Sie verstehen nicht, wie du auf deine Lösung kommst. Beschreibe und skizziere, welche Vergleiche und Beobachtungen du gemacht hast. Erkläre es so, dass Xenia & Xaver deinen Vergleich selber durchführen können.

Tab. 5.6 Standardisierte feeding back und feeding forward Angebote für Aufgaben zum effektbasierten Vergleichen.

Kompetenzen Die Schülerinnen und Schüler können...	**Feeding back**	**Feeding forward**
...einen theoretisch funktionierenden Vergleich durchführen.	☹ Dein Vergleich bei **A**1 funktioniert nicht oder man erkennt nicht, was du machst.	Mache einen anderen Vergleich bei **A**1 und beschreibe ihn sorgfältig. *oder* Hast du alles notiert? Überlege dir, welche Informationen fehlen.
	😐 Dein Vergleich bei **A**1 funktioniert, aber deine Beschreibungen oder Daten enthalten Fehler.	Achte darauf, beim Vergleich bei **A**1 immer denselben Effekt zu erzeugen. Achte darauf, dass deine Daten bei **A**1 nicht widersprüchlich sind.
	☺ Dein Vergleich bei **A**1 funktioniert und man versteht, was du machst.	Überlege dir, ob es sich lohnt, den Vergleich bei **A**1 zu wiederholen.
...eine qualitative Reihenfolge von drei Objekten aufstellen.	☹ Man erkennt nicht, dass du bei **A**3 alle drei Objekte in eine Reihenfolge gebracht hast.	Achte darauf, dass du bei **A**3 alle drei Objekte vergleichst.
	😐 Deine Objekte bei **A**3 stehen in einer Reihenfolge, aber man versteht nicht, wie du darauf gekommen bist.	Achte darauf, dass deine Daten bei **A**3 vollständig sind. *oder* Achte darauf, dass deine Daten bei **A**3 nicht widersprüchlich sind.
	☺ Du hast eine Reihenfolge bei **A**3 aufgestellt und man sieht, wie du darauf gekommen bist.	Überlege dir, ob es sich lohnt, den Vergleich bei **A**3 zu wiederholen.
... Bedingungen für einen fairen Vergleich beschreiben und anwenden.	☹ Deine Vergleiche sind nicht fair.	Achte darauf, dass du Gleiches mit Gleichem vergleichst.
	😐 Deine Vergleiche sind noch nicht ganz fair.	Überlege dir, welche Bedingungen hier alle gleichbleiben müssen.
	☺ Du hast faire Vergleiche gemacht.	Überlege dir andere faire Vergleiche, die du hier machen könntest.
…Objekte mithilfe quantitativer Aussagen vergleichen.	☹ Dein Vergleich bei **A**4 funktioniert nicht oder man erkennt nicht, was du machst.	Mache einen anderen Vergleich bei **A**4 und beschreibe ihn sorgfältig. *oder* Achte darauf, dass du bei **A**4 alle drei Objekte miteinander vergleichst.
	😐 Dein Vergleich bei **A**4 funktioniert, aber es fehlen noch wichtige Beschreibungen oder Daten.	Achte darauf, dass deine Daten bei **A**4 vollständig sind. *oder* Achte darauf, dass deine Daten bei **A**4 nicht widersprüchlich sind.
	☺ Dein Vergleich bei **A**4 funktioniert und man versteht, was du machst.	Überlege dir, ob es sich lohnt, den Vergleich bei **A**4 zu wiederholen.

6 Validating performance assessments: measures that may help to evaluate students' expertise in 'doing science'

Hild, P., Gut, C. & Brückmann, M. (2018). Validating performance assessments: measures that may help to evaluate students' expertise in 'doing science'. *Research in Science & Technological Education*.

6.1 Abstract

Background. Several different measures have been proposed to solve persistent validity problems, such as high task-sampling variability, in the assessment of students' expertise in 'doing science'. Such measures include working with a priori progression models, using standardised item shells and rating manuals, augmenting the number of tasks per student, and comparing different measurement methods.

Purpose. The impact of these measures on instrument validity is examined here under three different aspects: structural validity, generalisability, and external validity.

Sample. Performance assessments were administered to 418 students (187 girls, ages 12–16) in grades 7, 8, and 9 in the two lowest school performance tracks in (lower) secondary school in the Swiss canton of Zurich.

Design and methods. Students worked with printed test sheets on which they were asked to report the outcomes of their investigations. In addition to the written protocols, direct observations and interviews were used as measurement methods. Evidene on the instruments' validity was reported by using different reliability and generalisability coefficients and by comparing our results to those found in literature.

Results. An a priori progression model was successfully used to improve the instrument's structural validity. The use of a standardised item shell and rating manual ensured reliable rating of the written protocols ($.79 \leq p_0 \leq .98$; $.56 \leq \kappa \leq .97$). Augmenting the number of tasks per student did not solve the challenge of reducing task-sampling variability. The observed performance differed from the performance assessed via the written protocols.

Conclusions. Students' performance in doing science can be reliably assessed with instruments that show good generalisability coefficients ($\rho^2 = 0.72$ in this case). Even after inplementing the different measures, task-sampling variability remains high ($\widehat{\sigma}^2_{pt} = 47.2\%$). More elaborate studies that focus on the substantive aspect of validity must be conducted to understand why students' expertise as shown in written protocols differs so markedly from their observed performance.

6.2 Introduction

The descriptions of expertise in what is known as ‘doing science’ (Hodson, 2009) that are found in educational goals and standards differ markedly from one country to another. These variations result from different historical and cultural backgrounds (DeBoer, 2000). A new Swiss national curriculum has been introduced that has various science-educational goals based on the results of a large-scale assessment (Labudde et al., 2012). In this assessment, paper and pencil tests, which typically neglect important components of doing science, were partly replaced by alternative ways of assessment based on students’ performance of concrete and meaningful investigations called ‘performance-based assessments’, or PAs (Shavelson, Baxter & Pine, 1991).

PAs involve samples of performances from several domains, with the resulting scores interpreted in terms of typical, or expected, performances in these domains (Kane, Crooks & Cohen, 1999). PAs have multiple advantages over paper and pencil tests; for example, both the process and the product of students’ performances may be examined (Schreiber, Theyssen & Schecker, 2016; Toh & Woolnough, 1990). In addition, PAs are believed to promote gender-neutral testing, to stimulate important skills such as the control-of-variables strategy, and to improve young students’ general performance, because all decisions are tied to actions (Berry, 1991; Jovanovic, Solano-Flores & Shavelson 1994; Gott & Duggan, 2002; Ruiz-Primo & Shavelson, 1996; Schwichow et al., 2016). Developing and in particular validating PAs is challenging (e.g. Stecher et al., 2000). The aim of this article is to discuss and investigate several measures that different researchers have proposed to improve the validity (and hence the quality) of PAs in science.

6.3 Validity of Performance Assessments

In order to assess the impact of several measures on validity, the term 'validity' needs to be specified. Based on the unitary view of construct validity postulated by Messick (1989; 1994; 1996), the validity of PAs may be defined as 'the evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of achievement assessment'. According to Messick (1996), six different aspects of validity may be investigated with PAs: *content*, *substantive*, *structural*, *generalisability*, *external*, and *consequential*.

In the last two decades, large-scale assessments using PAs in science have been administered using instruments that consist either (1) of several tasks from the same content domain, but requiring different conceptual and/or procedural knowledge (Gott & Duggan, 1996; Stecher et al., 2000); or (2) of a few tasks that require the same knowledge (i.e. the same type of investigation) but from different content domains (Shavelon & Ruiz-Primo, 1998).

One important threat to the structural validity of these PAs is their dependence on reading and writing (Haertel & Linn, 1996). Previous studies have also reported a lack of generalisability of the instrument's output (Shavelson, Ruiz-Primo & Wiley, 1999) in addition to a lack of evidence for structural (Gott & Duggan, 2002) and external (Ruiz-Primo & Shavelson, 1996) validity. These three aspects (structural validity, generalisability, and external validity) will be discussed in the following subsections. Table 1 shows several exemplary studies where one of these aspects is being focused.

6.3.1 Structural validity

The aspect of structural validity examines 'the scoring system as it relates to the construct domain' (Miller & Linn, 2000, 370). Messick (1996, 10) points out that what is known about the structural relations, inherent in behavioral manifestations of the construct in question, should be rationally consistent with a scoring model. In orther words, the theory of the construct domain should not only guide the selection or construction of relevant assessment tasks, but also the rational development of construct-based scoring criteria. According to Miller and Linn, multiple options may be used when scoring complex responses. Scoring is usually done with a rating manual using multiple raters, who either focus on pieces of the assessment or on the performance as a whole. To evaluate if the scoring system is consistent with the construct domain, classical testing and item-response theory (such as Rasch modelling) may be used. The scoring system itself may focus on knowledge, skills, or other

attributes that are theoretically determined prior to the testing using expert judgements (i.e. content validity).

Tab. 6.1 Exemplary studies and their focus (X) on construct validity.

	structural validity	**generalisability**	**external validity**
Cronbach et al. (1997)		**X**	
Gao, Shavelson & Baxter (1994)		**X**	
Gott & Duggan (2002)	**X**		
Gut (2012)	**X**		
Harmon et al. (1997)	**X**		
Hammann, Phan, Ehmer & Grimm (2008)			**X**
Ruiz-Primo & Shavelson (1996)			**X**
Schreiber, Theyßen & Schecker (2014)			**X**
Shavelson, Ruiz-Primo & Wiley (1999)		**X**	
Vorholzer, von Aufschnaiter & Kirschner (2016)	**X**		**X**
Webb, Schlackman & Sugrue (2000)		**X**	**X**

Different progression dimensions may be evaluated to judge whether someone has become more literate in doing science. According to Gut et al. (2014), students may improve in several ways. They may:

- solve similar problems with higher complexity (problem complexity);
- work more independently (autonomy);
- solve similar tasks in multiple contexts (transfer ability);
- show more stable results (performance stability).

In addition, their problem solutions may increase in quality (solution quality).

6.3.2 Generalisability

A second validity aspect is the instrument's general application in different settings. According to Messick (1996, 7), the generalisability aspect 'examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including validity generalization of test-criterion relationships'. Different sources of measurement errors ('facets') may be investigated within generalisability (G) studies (Brennan, 1996; Cronbach, Rajaratnam & Gleser, 1963). In these studies, 'the replicability or consistency of assessment results across multiple levels of random facets of an assessment is examined to understand the boundaries of the construct' (Miller & Linn, 2000, 370). G studies divide the total variance of scores $\hat{\sigma}_X^2$ into the variances of a study's different sources (*s*) by calculating variance components $\hat{\sigma}_s^2$ (Shavelson & Webb, 1981). Based on the relative weight of the variance components, a generalisability coefficient $\hat{\rho}^2$ is computed, which may be interpreted as an indicator of an instrument's reliability.

Most studies include raters and tasks as two major sources of error. In our G studies, tasks represent experimental problems to be solved. Other facets often included in G studies include 'occasions' (Webb, Schlackman & Sugrue, 2000); measurement methods (Shavelson, Gao & Baxter, 1993); and classes, schools, or regions (Cronbach et al., 1997; Gao et al., 1994; Taut & Rakoczy, 2016). The students themselves are not treated as facets in G studies, but they do constitute the population (Brennan, 2000).

6.3.2.1 Tasks

In their study, Gao et al. (1994) found a strong student-by-task interaction as well as a moderate school-by-task interaction; Webb, Schlackman, and Sugrue (2000) reported similar results in their G study. No set of rules appears to guarantee that tasks that are similar on the surface will elicit similar performances from students (Stecher et al. 2000). In all Pas known to the authors (e.g. Gao et al., 1994; Shavelson et al., 1999; Webb, Schlackman & Sugrue, 2000), task-sampling variability has led to the highest variance component among G studies, thus rendering it impossible to use these assessments for large-scale examinations, where student performance should be assessed with a small number of tasks.

6.3.2.2 Raters

In crossed study designs that have included people, tasks, and raters (p x t x r), all variance components in which the study investigated rater influence, namely $\hat{\sigma}^2_{pr}, \hat{\sigma}^2_{tr},$ and $\hat{\sigma}^2_{ptr}$, were found to be low when inter-rater reliability was high (Brennan, 1996; Shavelson, Gao & Baxter, 1993). This finding means that in PAs that show high inter-rater reliability, raters could easily be excluded from G studies. In other words, high inter-rater reliability may be viewed as a 'necessary, although not sufficient condition' for generalising a performance test (Brennan, 2000, 346).

6.3.2.3 Occasions

At different occasions, the same or very similar tasks may be solved at different times. From a classical perspective, sampling variability due to the existence of different occasions thus most closely corresponds to the notion of test-retest reliability (Brennan, 1996). According to several studies' results in the literature, the variance that is attributable to the interaction of people, tasks, and occasions may be very large (Ruiz-Primo, Baxter & Shavelson, 1993; Webb, Schlackman & Sugrue, 2000).

6.3.3 External Validity

Studies on external validity analyse 'the relationship of test scores to variables external to the test … Patterns of high, moderate, and low empirical relationships are examined, based on theoretical expectations or hypotheses' (Miller & Linn, 2000, 372). The subsections below present several threats to external validity.

6.3.3.1 Different Measurement Methods

The process and products of student performance may be measured with direct observations, concept maps, written protocols (i.e. notebooks), and computer log files, among other methods. Although different measurement methods may also be used in G studies, most researchers have treated this aspect while examining the external validity of a certain instrument (Shavelson, Gao & Baxter, 1993). Written protocols may be considered as a valid measure of student performance, if their outcome correlates highly with the outcome assessed through direct observations. In Shavelson, Ruiz-Primo, and Wiley's (1999) study, direct observations were found to correlate more highly with written protocols than with computer simulations and short-answer testing. But method heterogeneity is an issue in many studies

on PAs (Ruiz-Primo & Shavelson, 1996; Schreiber et al., 2014). For example, Gott and Duggan (2002) found a weak correlation between direct observations and written protocols in their study. Hamman and colleagues (2008) even stated that PAs and multiple-choice tests seem to rely on different constructs.

6.3.3.2 Different Variables

Vorholzer, von Aufschnaiter & Kirschner (2016), who investigated the relationship between cognitive learner variables and student performance in PAs, found a model-person reliability of 0.80. The importance of certain affective variables may also be examined when dealing with scientific PAs. Subject- and topic-related situational interest, intrinsic and extrinsic motivation (Fechner, 2009), self-concept in natural science (Marsh et al., 2005), and the expectancy of self-efficacy (Pintrich & De Groot, 1990) may positively influence the cognitive outcome of a PA. Based on Hofstein's (2004) review, a student's interest in conducting experiments could also have a positive impact.

6.3.3.3 Fairness

A further aspect of external validity is fairness (Haertel & Linn, 1996). PAs should not provide one subpopulation (gender, language, social background, or race, for example) with an advantage over other subpopulations. One important finding on this topic is that gender differences are thought to be less dominant (or even negligible) in PAs than with other measurement formats, such as multiple-choice tests (Jovanovic, Solano-Flores & Shavelson, 1994).

6.3.4 Different measures to improve validity

Researchers have proposed a variety of different measures to improve the persistent validity problems when using PAs.

First, working with a priori progression models may help to elicit evidence of the instrument's structural validity (Gut, 2012; Gut et al., 2014).

Second, to ensure good reliability of the instrument, tasks and assessments should be standardised as much as possible (Kane, Crooks & Cohen, 1999; Solano-Flores et al., 1999; Stecher et al., 2000). Solano-Flores et al. (1997) proposed working with standardised item shells when designing PAs. These shells are helpful in standardising the format and level of inquiry of the tasks, thus keeping construct-irrelevant difficulties constant, and they should

help to avoid undue variance in student performance, thereby producing more reliable and generalisable assessments (Solano-Flores et al., 1999). Previous studies have reported, however, that these shells do not help in solving the task-sampling variability problem, nor can they ensure the interchangeability of measurement methods (Shavelson, Ruiz-Primo & Wiley, 1999; Webb, Schlackman & Sugrue, 2000; Stecher et al., 2000).

Third, in order to improve the generalisability coefficient in G studies, researchers have proposed the augmenting of the number of tasks per student as a further method. Based on decision-study considerations, the number of tasks required for generalisability coefficients higher than .70, which result in acceptably small error variance and/or acceptably large reliability-like coefficients, is two to ten (Miller, 1998). Shavelson, Gao, and Baxter (1993) recommend 10 tasks, each solved in 15 minutes, in order to obtain generalisability coefficients higher than .80.

Fourth, assumptions about the external validity (and generalisability) of PAs should be made only after comparing different assessment and measurement methods (Webb, Schlackman & Sugrue, 2000).

The following section uses various research questions to focus on the impact of these four measures under the three aspects of validity (structural validity, generalisability, and external validity) where an improvement is expected.

6.4 Research Questions

Most of the studies mentioned above assessed expertise in ‘doing science’ using (1) a large number of students; (2) small numbers of tasks, problems, and occasions; (3) one or two measurement methods; and (4) one aspect of validity. These studies have reported evidence on structural validity by examining the consistency of test scores with the constructs under investigation using post-hoc progression modelling (Gut, 2012). For generalising the results of PAs, high inter-rater reliability values have been shown to ensure negligible variance because of raters; studies have also found that student performance mostly depends on the sampled problems and the measurement methods that are used.

Ruiz-Primo and Shavelson (1996) have reported that not all assessment and measurement methods seem to be interchangeable. Studies on external validity typically focus on different measurement methods, the influence of different learner variables, or the fairness of the instrument. A more systematic investigation that would compare several measurement methods, treat multiple aspects of validity, and include the proposed measures is currently missing in the literature, which has led us to create the following research questions.

RQ1: Is the instrument valid in the sense of structural validity, generalisability, and external validity?

RQ2: To what extent do the four proposed measures contribute to the instruments’ validity.

RQ2.1: Does the structure and progression found with the scoring system match the theoretical assumptions based on an a priori model (structural validity)?

RQ2.2: Can the standardised rating manual be used for direct observations and written protocols?

RQ2.3: Does augmenting the number of tasks per student, lead to higher generalisability coefficients as predicted by Miller (1998) and Shavelson and colleagues (1993) (generalisability)?

RQ2.4: How is actual student performance (based on direct observations and interviews) related to the test scores from their written protocols (external validity)?

6.5 The Instrument

In our research, we work with an instrument that is used to evaluate a person's expertise in 'doing science', such as observing, making measurements, or investigating (Metzger et al., 2014b). This instrument consists of around thirty different experimental problems and was developed for students in lower secondary Swiss schools (ages 12–16) in poorly performing school tracks. The studies presented below focus on one exemplary form of doing science, namely conducting comparative investigations (CIs).

CIs, which may be thought of as an example domain for performance assessments (Shavelson et al., 1998), may be defined as tasks where two or more objects are compared in terms of a given attribute while other variables are controlled (Solano-Flores & Shavelson 1997). Past studies have used CIs in large-scale assessments (Erickson, 1994; Harmon et al., 1997; Stecher, 1996). Except for Solano-Flores's (1994) and Shavelson, Solano-Flores, and Ruiz-Primo's (1998) studies, only a few have focussed on the validation of CIs (Hungerford & Miles, 1969; Meyer & Carlisle, 1996; Tomera, 1974).

Eight different experimental problems were designed using an a priori progression model for conducting comparative investigations (Gut et al., 2017; Hild et al., 2015). The model consists of four content and performance standards (including cognitive processes), as well as ideas about the nature of scientific comparisons, called quality standards (QSs), of the students' solutions (see Figure 6.1).

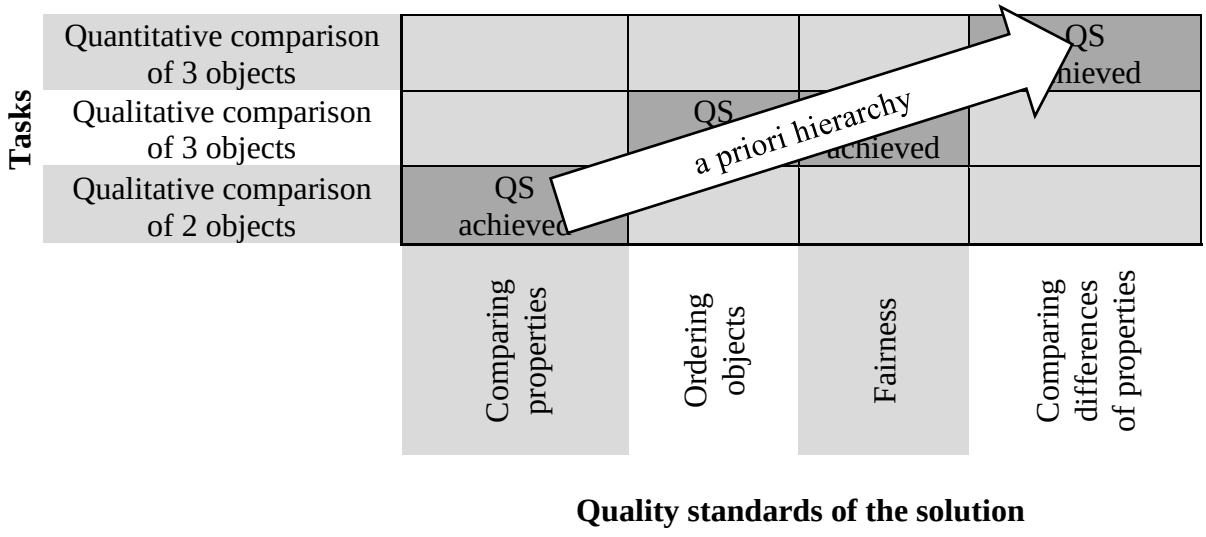


Fig. 6.1 A priori progression model for comparative investigations (CIs).

In all eight problems, students were asked to compare certain property variables such as acidity, solubility, water quantity, energy efficiency, the osmotic effect, or the magnetic strength of different objects using qualitative (i.e. defining sequences from strongest to weakest) and quantitative (i.e. determining which objects are more similar) descriptions (Hild, Metzger & Parchmann, 2018). The content domains of these problems were generated from authentic biological, chemical, and physical contexts (Table 6.2). Some of the problems were taken from national or international large-scale assessments, such as the Assessment of Performance Unit (APU), the Trends in International Mathematics and Science Study (TIMSS), and Harmonization of Compulsory Education (HarmoS). Others were derived from PAs developed at the Stanford Education Assessment Laboratory (SEAL) in the 1990s by Richard Shavelson and colleagues or were developed at our university.

Tab. 6.2 Content domains in the comparative investigations (CIs).

title	**students are asked to compare…**
Juices	*the amount of water in different fruits (cucumbers, apples, and zucchini).*
Dairy products	*the water properties of different stabilised water-in-oil and oil-in-water emulsions (curd, soft cheese, body lotions, and hand creams).*
Magnets	*the magnetic strength of different magnets.*
Potatoes	*the osmotic effect of different powders (salt, sugar, flour) on potatoes.*
Powders	*the solubility of different powders (monosaccharide, disaccharide, salt) in water.*
Solar cells	*the efficiency of different solar cells.*
Nuts	*the amount of fat in different nuts (almonds, hazelnuts, and walnuts).*
Teas	*the acidity of different herbal infusions (rooibos, fruit tea, chamomile) dissolved in water.*

All problems were formatted using a standardised item shell (Table 6.3), and student expertise was assessed using a standardised rating manual (Ruiz-Primo, Baxter & Shavelson, 1993; Solano-Flores et al., 1997). The basic problem (task 1) within the eight CIs was to invent and accomplish an adequate experimental setting for comparing two objects with respect to a given property variable. Further tasks that were present in every CI involved ordering three objects (task 2), judging the fairness of a comparison (task 3), and identifying the two objects whose properties were most similar (task 4). One example of an experimental problem has

been translated and included at the end of this book (‘Dairy products’). All materials that were used to solve these problems may be handled without further safety instruction and may be found in standard school labs in Switzerland.

Table 6.3 Item shell for designing CIs.

Title	[*One or two words such as Powders, Dairy products, or Magnets*]
material	[*Picture and terms of all material needed for the comparative investigation*]
problem	[*Girl’s name*] and [*boy’s name*] must find out which [*object*] has the best [*property variable*]. They don’t know how to do it. Solve the problem for [*girl’s name*] and [*boy’s name*] by doing the following three tasks.
task 1	Compare [*object*] ***A*** with [*object*] ***B***. Find out which [*object*] has the best [*property variable*]. What did you find out? Mark the correct answer: o [*Object*] **A** has a better [*property variable*] than [*object*] **B**. o [*Object*] **B** has a better [*property variable*] than [*object*] **A**. [*Girl’s name*] and [*boy’s name*] have found a different result. They want to know how you managed to find yours. Describe and sketch which experiments and observations you did. Explain this to [*girl’s name*] and [*boy’s name*] in such a way that they will be able to do the comparative investigation themselves.
task 2	Continue by including [*object*] ***C***. Find out a sequence for [*objects*] **A**, **B**, and **C**. Start with the [*object*] that has the best [*property variable*]. [*Girl’s name*] and [*boy’s name*] have found a different sequence. They don’t understand how you found your sequence. Describe and sketch which experiments and observations you did. Explain the steps to [*girl’s name*] and [*boy’s name*] in such a way that they will be able to do the comparative investigation themselves.
task 3	How did you ensure that your comparative investigations would be fair?
task 4	Find out which two [*objects*] are most similar, if you examine the [*property variable*]. If necessary, do further comparisons to solve this task. What did you find out? Mark the correct answer: o [*Object*] **A** and [*object*] **B** are most similar. o [*Object*] **A** and [*object*] **C** are most similar. o [*Object*] **B** and [*object*] **C** are most similar.

Similarly to Commons et al.'s (2008) study, a postulated hierarchy among these standards could be partly validated using item-response theory (i.e. Rasch-scaled stage scores). The first analysis showed good fit values for the two samples of 152 and 190 students (see Table 6.4, from Gut et al., 2017). To answer RQ2.1, more emphasis must be placed on the progression along the quality standards using classical testing.

Tab. 6.4 Characteristic values of unidimensional Rasch testing.

	sample A	**sample B**
size	152	190
tasks * quality standards	24	24
interrater reliability p_0	> .79	≥ .79
interrater correlation κ	> .61	> .58
infit, T values	.93–1.09; $\lvert T \rvert \leq .9$	0.97–1.20; $\lvert T \rvert \leq 1.4$
outfit, T values	$\lvert$outfit-1$\rvert < 0.15; \lvert T \rvert \leq 1.1$	$\lvert$outfit-1$\rvert < 0.25; \lvert T \rvert < 2$
discrimination values	.69 – .80	.34 – .63
item separation reliability	.98	.98
EAP/PV-reliability	.65	.65
variance	0.46	0.53

For the generalisability of our instrument (RQ2.3), an initial G study was conducted using a sample of 190 students based on only two different CIs at two different occasions (Hild et al., 2018b). A very high student-task interaction (around 70% variance) was found in this study. No research on the external validity of the instrument had yet been undertaken at the time this study was conducted.

6.6 Assessing the impact of proposed measures on the instruments' validity

Table 6.5 shows the different studies on the impact of the proposed measures. The studies were carried out with 418 students (187 girls; 44.7%, ages 12–16) from grades 7, 8, and 9 in the two lowest school-performance tracks (called A and BC) in lower secondary school of the canton of Zurich. A is the second-highest track in Swiss secondary school. Data was recorded through school testing (samples A–D) using written protocols and in cognitive labs (sample D) using written protocols, direct observations, and interviews (Table 6.6). Students from samples A, B, and C solved two out of eight experimental problems on their own. Students from sample D solved four problems (two in school testing on their own and two in cognitive labs in pairs). Permission to publish video and audio material from the students (sample D) was obtained.

6.6.1 Assessing structural validity

In this section, we elicit evidence on the instruments' structural validity, with the hope of answering RQ1 (structural aspect) as well as RQ2.1: Does the structure and progression found with the scoring system match the theoretical assumptions based on an a priori model?

Tab. 6.5 Information on sample size used in the different validity studies.

aspects of validity	**structural validity**	
	i. consistency of the scoring system (RQ2.1)	**418** students (samples A–C)
	Generalizability	
	ii.1. generalisability of the rating manual (RQ2.2)	**418** students (samples A–D)
	ii.2. generalisability of the test scores (RQ2.3)	**342** students (samples A–B)
	external validity	
	iii. different measurement methods (RQ2.4)	**16** students (sample D)

Tab. 6.6 Information on assessment and measurement methods used in the validity studies.

	assessment methods	**school testing**	**cognitive labs**	
	measurement methods	written protocols	written protocols and direct observations	interviews
samples	**sample A (2 problems / 2 occasions)** Juices, Dairy Products and Magnets	**152** students (63 girls) 7BC (26) 7A (44) 8BC (17) 9BC (24) 9A (41)		
	sample B (2 problems / 2 occasions) Potatoes, Powders and Solar Cells	**190** students (94 girls) 7BC (54) 7A (63) 9BC (23) 9A (50)		
	sample C (2 problems / 1 occasion) Potatoes, Powders and Dairy Products	**76** students (30 girls): grade 9BC (40) and grade 9A (36)		
	sample D (4 problems / 2 occasions) content domains from sample C + Nuts, Teas and Juices	**16** students from sample C (13 girls): grade 9BC (8) and grade 9A (8)		

6.6.1.1 Methods

Sample. In this study, 418 students solved two CIs, each in 18 minutes on one occasion (sample C) or two occasions (samples A and B); see Table 6.6. Students worked with printed test sheets on which they were asked to write down their answers and to report the outcomes of their investigations. Figure 6.2 shows a school testing situation. Even when students were seated next to each other, they all solved different experimental problems.

Fig. 6.2 School testing.

Pilot testing. The eight CIs were designed in an iterative process by four different assessment developers working with a standardised item shell (Table 6.3). All CIs were piloted with at least 60 students from lower secondary school; one assessment developer was always present in the classroom to make observations and to collect student and teacher feedback on the content and complexity of the different CIs. Table 6.4 shows initial evidence for structural validity when working with Rasch-scaled stage scores (and EAP/PV reliabilities of .65).

Rating manual. A standardised rating manual was developed and constantly optimised during the different raters' pilot coding (Table 6.7). Special rater training was used to ensure high inter-rater reliability. The manual contained content-specific cues and explanations of how the different criteria should be coded. Every CI was rated by at least two people and as

one single item. The four quality standards (QSs) of the students' solutions were examined in the assessment by clustering up to three criteria at a time. Every QS was accomplished when two-thirds of the criteria were rated 1. The test scores were calculated by summing the QSs that were reached in each CI.

Tab. 6.7 Overview of the four quality standards (QSs) used in the rating manual.

QS	**description (criteria)**
I. Comparing properties	*Setting up an adequate experiment; drawing a correct conclusion from data (3)*
II. Ordering objects	*Drawing the correct order (conclusion) from the data (3)*
III. Fairness	*Naming at least two conditions for a fair comparison (1)*
IV. Comparing differences of properties	*Setting up an adequate experiment; drawing the correct conclusion from the data (3)*

6.6.1.2 Results

The inter-rater reliability values of the written protocols were good ($.79 \leq p_0 \leq .98$; $.56 \leq \kappa \leq .97$). As Figure 6.3 shows, the a priori hierarchy for samples A, B, and C was always exposed. The QSs significantly differed in 26 of 27 cases (the stars correspond to degrees of significance in the Wilcoxon test). The last QS, which compared differences of properties (not indicated here), was always the lowest and was almost never achieved. This QS differed significantly from all other QSs in the Wilcoxon testing. The frequencies of the solved QSs (indicated in percentages) differed markedly between the three samples: 81% of all students from sample A (with a smaller percentage of students in the 7th grade) reached the first QS. When two CIs ('potatoes' and 'powders') were administered to students in samples B and C, the frequencies of solving the QSs differed, but the hierarchy persisted.

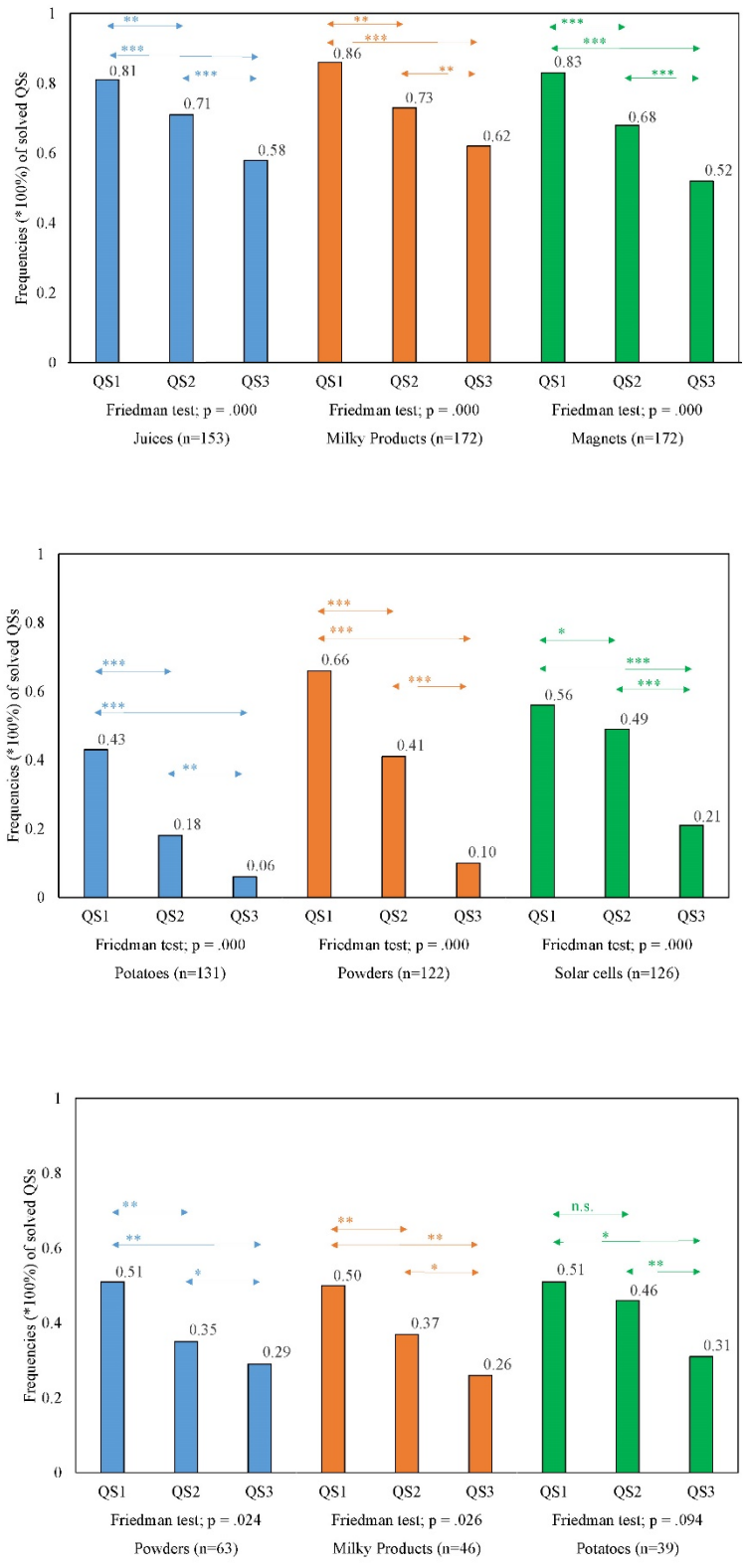


Fig 6.3 Frequencies (*100%) of solved quality standards (QSs) in different CIs for samples A, B, and C.

6.6.1.3 Discussion

The structural aspect of RQ1 and RQ2.1 can be answered with a yes under the restriction that the fourth quality standard is beeing removed. Designing experimental problems based on (and assessing) students' expertise with an a priori progression model assured a solid relationship between the scoring system and the construct domain. Test scores from all CIs showed that the postulated a priori hierarchy among the first three QSs existed in every single CI when examining the written protocols. Because the fourth QS was almost never achieved, the progression model should be reduced to three standards when working with students of these ages and tracks. One reason that the fourth QS was almost never achieved could be that many young students stop investigating as soon as they achieve a desired or interesting outcome (Millar et al., 1996; Schauble, Klopfer & Raghavan, 1991), and they avoid a more systematic testing of variables.

In sample C, the differences between standards were less significant (including one non-significant difference) than in A or B. This outcome could have been due to the lower number of people (n = 76) who were solving the CIs. But students from sample C were also expected to reach higher frequencies of solved QSs for every CI because of the lack of students from the 7th and 8th grades in that sample (Table 6.6). One reason could be that the raters applied the rating manual differently for each sample. Because the eight types of content used in the CIs were unknown to the students, the progressions along the standards seemed to have been independent of the content domain (Table 6.2). If similar problems were administered to students using known contents, then their engagement in solving the tasks could increase, which could in turn lead to higher test scores (Hart et al., 2000). Finally, the QSs may be of great use to foster students' expertise in conducting CIs, because the data can still be interpreted, thanks to the data's descriptive nature.

6.6.2 Assessing Generalisability

In this section, we answer RG1 (generalisability aspect), RQ2.2, and RQ2.3.

RQ2.2: Can the standardised rating manual be used for direct observations and written protocols?

RQ2.3: Does augmenting the number of tasks per student, lead to higher generalisability coefficients as predicted by Miller (1998) and Shavelson and colleagues (1993)?

6.6.2.1 Methods

Sample. RQ2.2 was examined by comparing the inter-rater reliability of the test scores from written protocols (samples A–D) and direct observations (sample D). The same rating manual has been used for both methods. To answer RQ2.3, a G study was conducted using the written protocols of all students from samples A and B. In both samples, two out of three problems had to be solved (Table 6.6).

G studies. Estimation of the different variance components of our G study design was performed with SPSS 22.0 (using the VARCOMP procedure) and the G2 programme (Mushquash & O'Connor, 2006). The study included tasks/problems and occasions as facets in a crossed design (p x t x o) with a total variance composed of seven facets (see Table 6.8):

$$\hat{\sigma}^2_{X_{pto}} = \hat{\sigma}^2_p + \hat{\sigma}^2_t + \hat{\sigma}^2_o + \hat{\sigma}^2_{pt} + \hat{\sigma}^2_{po} + \hat{\sigma}^2_{to} + \hat{\sigma}^2_{pto,e}$$

Tab. 6.8 Interpretation of variance components for the generalisability (G) study.

Variance component	**interpretation**
person	*systematic differences between people. Tasks and occasions overlap.*
task/problem	*systematic inconsistencies between tasks – independent of people and occasions.*
occasion	*systematic inconsistencies between occasions – independent of people and tasks.*
person x task	*judgements about tasks depend on people, independent of occasions.*
person x occasion	*judgements about occasions depend on people, independent of tasks.*
task x occasion	*judgements about occasions depend on tasks, independent of people.*
person x task x occ.	*unspecified error variance; three-way interaction confounded with residual.*

In our study, students solved two very similar problems in the same week. The design was not fully crossed, as each student solved two out of three experimental problems and because two different samples with different problems (six problems overall) were included in the study (see Table 6.6). Hence, the term $\hat{\sigma}^2_{pto,e}$ is zero. All facets were considered random (Gao et al., 1994; Shavelson & Webb, 1981). The 'minimum norm quadratic unbiased estimation' (MINQUE) technique was chosen to optimise variance and covariance components (Webb,

Shavelson & Haertel, 2006). This estimator does not assume normality and does not involve an iterative estimation. The weights corresponding to the relative sites of the variance components were placed a priori to zero. We then calculated the percentages attributed to each source of error over the total variance, as well as the generalisability coefficients for relative decisions.

6.6.2.2 Results

This section presents the results related to RQ2.2 and RQ2.3.

RQ2.2: Can the standardised rating manual be used for direct observations and written protocols?

As Table 6.9 shows, the inter-rater reliability values for both the rating of written protocols and direct observations were high enough to support a generalisation of the manual. Student performance could be assessed independently of the measurement method used with the same rating manual, which does not mean that the students reached equal numbers of (nor the same) QSs when both measurement methods were used.

Tab. 6.9 Interrater reliabilities of written protocols and direct observations.

	written protocols (sample A–D)	**direct observation (sample D)**
interrater reliability	$.79 \leq p_0 \leq .98$	$.81 < p_0 \leq 1$
interrater correlation	$.56 \leq \kappa \leq .97$	$.62 < \kappa \leq 1$

RQ2.3: Does augmenting the number of tasks per student, lead to higher generalisability coefficients as predicted by Miller (1998) and Shavelson and colleagues (1993)?

Table 6.10 shows the estimated variance components in the crossed design (p x t x o). The largest contributors to total variance were people and person-task interactions. The largest of all variance components was $\hat{\sigma}^2_{pt}$, which suggests a considerably different rank ordering among the students' mean test scores (the sum of QSs was between 0 and 4) for each CI. The variance attributable to the interaction of people or tasks with occasions was very small.

Tab. 6.10 Estimated variance components in the p x t x o design.

source	**variance component**	**estimate**	**percent of total variability**
person (p)	$\hat{\sigma}_p^2$	0.500	20.5
task (t)	$\hat{\sigma}_t^2$	0.380	15.6
occasion (o)	$\hat{\sigma}_o^2$	0.293	12.0
p x t	$\hat{\sigma}_{pt}^2$	1.151	47.2
p x o	$\hat{\sigma}_{po}^2$	0.000	0
t x o	$\hat{\sigma}_{to}^2$	0.113	4.6
p x t x o,s*e*	$\hat{\sigma}_{pto,e}^2$	0.000	0
relative error variance	$\hat{\sigma}_\delta^2$	0.192	
absolute error variance	$\hat{\sigma}_\Delta^2$	0.411	
generalizability coefficient	$\hat{\rho}^2$	0.723	
dependability coefficient	$\hat{\Phi}^2$	0.549	

6.6.2.3 Discussion

As this study showed, inter-rater reliability levels of κ>.60 allowed for an effective use of the rating manual for both written protocols and direct student observations, which is a prerequisite for answering RQ2.4 in the following study. Direct observations were rated slightly more reliably than ratings for written protocols. Possible reasons that the reliability was lower with written protocols were that many of the students from lower tracks had difficulties expressing themselves, or that students forgot important information when writing down the different steps they did while comparing the objects.

In addition to the raters, other sources of measurement error were investigated using a G study to answer RQ2.3. With a generalisability coefficient $\hat{\rho}^2$ of 0.72, using six tasks and two occasions, our instrument showed even better results than predicted (Gao et al., 1994; Webb, Schlackman, and Sugrue 2000). Working with standardised item shells and careful rater trainings may explain this improvement. Sources of measurement error that explain parts of the total variance includes people ($\hat{\sigma}_p^2 = 20.5\%$), tasks ($\hat{\sigma}_t^2 = 15.6$), and occasions ($\hat{\sigma}_o^2 = 12\%$). While the third item may be explained by a training effect, judgements about occasions

were independent of people ($\hat{\sigma}^2_{po} = 0\%$). Hence, the role of occasions appeared to be negligible, and the tasks were seemingly interchangeable in time.

In contrast, person-task interactions ($\hat{\sigma}^2_{pt} = 47.2\%$) , which were high, represented the major source of measurement error. Augmenting the number of experimental problems from six (two per person) to a larger number, such as four per person, could reduce this amount and render the instrument even more reliable. The question remains, however, of whether the instrument could still be usable and practical in schools for large-scale assessments (Solano-Flores, 1999).

6.6.3 Assessing External validity

RQ2.4: How is actual student performance (based on direct observations and interviews) related to the test scores from their written protocols?

6.6.3.1 Methods

Sample. Eight student pairs (sample D) were invited to our university and tested in cognitive labs using written protocols, direct observations, and interviews (tables 6.5 and 6.6). These students had already solved two CIs in school testing.

Video coding. A camera was used to record the eight student pairs, who solved two CIs out loud. A wireless microphone was placed next to the table to reduce noise (Figure 6.4). All students received a five-minute training in thinking aloud before they started solving the CIs. Two people rated the video data using the standardised rating manual described above. The inter-rater reliability values of the written protocols and direct observations (videos) were measured using Cohen's kappa. Figure 6.5 shows a student pair working in a cognitive lab.

Interviews. The eight student pairs were interviewed immediately after solving the two CIs. In the interview, they retrospectively described how they had solved the problems and clarified their responses (Table 6.11). The QSs used to rate both the videos and written protocols were adopted in the interviews through task-specific questions.

Fig. 6.4 A cognitive lab with a camera and a wireless microphone.

Fig. 6.5 Two students working in pairs on the problem ‘Juices’.

Tab. 6.11 Interview with general and task-specific questions about the investigated CIs.

i) General questions about the CIs		
Did you like doing this investigation?		
	If yes…	What exactly did you like about it?
	If no…	What didn't you like about it?
What were you supposed to find out?		
Did you understand the problem and the different tasks?		
	If no…	What didn't you understand?
Do you know what [*property variable*] means?		
	If yes…	Can you explain it? Can you name examples of compounds or situations where the [*property variable*] is being investigated or indicated?
	If no…	What could this phrase mean, do you think?
Has the material been adequate for doing the comparative investigation?		
Are there any other instruments you would have liked to use for this investigation?		
Were you able to solve the four tasks?		
	If no…	Which task was difficult to solve, and why? What would have helped?
	If yes…	Which task was the most difficult, and why? What would have helped?
ii) Task-specific questions about the CIs		
Task 1	*If the task is solved incorrectly…*	What is the correct answer? Why?
	If contradictory findings are reported…	You wrote down that ... and that ... Which answer is the correct one? Why?
Task 2	*If no sequence of objects is reported, or if it is not clear what the student means…*	How did you determine this sequence?
Task 3	*If the student does not write down an answer…*	When would this comparative investigation be fair? What else should be considered?
	If the student only writes down one answer…	What else should be considered when doing a fair comparison?
Task 4	*If the student does not write down an answer…*	Which objects differ the least when comparing [*property variable*]? Why?
	If the student provides an answer…	How did you determine this result?
	…but the student reports contradictory findings…	You wrote down that ... and that ... Which answer is the correct one? Why?

6.6.3.2 Results

Direct observations vs written protocols. Sixteen CIs were videotaped, and the outcomes (written protocols and direct observations) were rated by two pairs of raters. Test scores were not attributed to single people but to the performance of the corresponding student pairs. One person rated both formats (see Table 6.12). Except for student pair 5 at occasion 1, the test scores resulting from direct observations were always higher or equal to the scores based on written protocols.

Tab. 6.12 Test scores and reached QSs from written protocols and direct observations.

student pair	occasion	score video	score protocol	Δ(difference)	**QS1** video	**QS1** protocol	**ΔQS1**	**QS2** video	**QS2** protocol	**ΔQS2**	**QS3** video	**QS3** protocol	**ΔQS3**	**QS4** video	**QS4** protocol	**Δ QS4**
1	1	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
	2	3	3	**0**	1	1	0	1	1	0	0	0	0	1	1	0
2	1	3	0	**3**	1	0	1	1	0	1	0	0	0	1	0	1
	2	3	0	**3**	1	0	1	1	0	1	0	0	0	1	0	1
3	1	3	0	**3**	1	0	1	1	0	1	0	0	0	1	0	1
	2	3	3	**0**	1	1	0	1	1	0	0	0	0	1	1	0
4	1	4	0	**4**	1	0	1	1	0	1	1	0	1	1	0	1
	2	4	0	**4**	1	0	1	1	0	1	1	0	1	1	0	1
5	1	3	4	**-1**	1	1	0	1	1	0	0	1	-1	1	1	0
	2	4	4	**0**	1	1	0	1	1	0	1	1	0	1	1	0
6	1	3	3	**0**	1	1	0	1	1	0	0	0	0	1	1	0
	2	3	0	**3**	1	0	1	1	0	1	0	0	0	1	0	1
7	1	4	3	**1**	1	1	0	1	1	0	1	0	1	1	1	0
	2	3	3	**0**	1	1	0	1	1	0	0	0	0	1	1	0
8	1	2	2	**0**	1	1	0	1	1	0	0	0	0	0	0	0
	2	4	3	**1**	1	1	0	1	1	0	1	0	1	1	1	0

Direct observations plus interviews. Interviews were only rated by one person and were solely used to understand why certain tasks (i.e. criteria) were answered incorrectly or incompletely. The students' answers to task-specific questions matched the outcomes of the direct observations. As the extracts from the interviews (Table 6.13) show, the student pairs who performed poorly did indeed have difficulties in correctly answering the questions in the interview. For example, to ensure a fair comparison and to achieve the third QS, according to the rating manual, students had to show (in the direct observations) and name (in the

interview) at least two conditions that would be held constant, which was not explicitly asked in the item shell. Student pairs who kept two conditions constant (for example, student pair 5) also named these two conditions in the interview. Those who did not keep two conditions constant could not name a second condition in the interview. (See interviews with student pair 1 or 8.)

Tab. 6.13 Extracts of interviews from different student pairs (*translated version*).

Interview with student pair 1; occasion 1 (*Nuts)*
Interviewer (I): Which objects differ the least when comparing the amount of fat? How did you find out this result? Student A: *I just observed which spot was the biggest and which spot was the second biggest.* I: OK, but it could have happened that one kind of nut leaves a big spot and the two other kinds of nuts leave tiny spots. What would you have done in that case? Would you have chosen the two tiny ones, or something else? Student A: *I don't understand; sorry.* I: Say you have three papers. Suppose that the spot on one is big, and the spots on the other two papers are very small. Which sorts of nuts would be more similar? Student B: *The two.* Student A: *The two big ones. If there were three nuts and two big ones, those would be more similar…*
Interview with student pair 5; occasion 2 (*Teas)*
I: When is this comparative investigation fair? What should be considered? Student B: *Measuring the time, we simultaneously put all the different teas in the cups and after 3 minutes measured the values. We also made sure that the water levels were equally high…*
Interview with student pair 8; occasion 1 (*Juices*)
I: When is this comparative investigation fair? You said weight. What else should be considered? Student B: *Maybe the size.* Student A: *The size?* Student B: *No, not necessarily, for example…it's complicated.* Student A: *No idea.* I: In the answer to which objects differ the least when comparing the amount of water, you said zucchini and cucumber. Why? Student B: *Both are vegetables.* Student A: *And they look the same.* Student B: *Yes, they look the same, and both are green. Apples are a kind of fruit.*

Figure 6.6 shows the inter- and intra-rater reliability values of (and between) the different measurement methods; because the interviews were only rated by one person, inter-rater reliability values are not included in the table. Intra-rater reliabilities (for sample D) could be calculated between written protocols and direct observations of student pairs. Cohen's κ was below .15, and the rater agreement never exceeded .68.

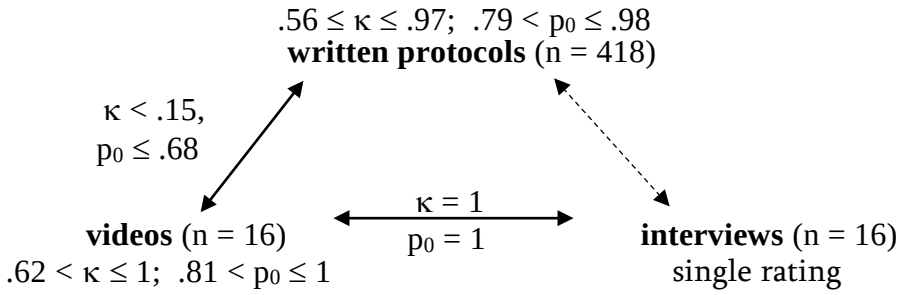


Fig. 6.6 Interchangeability of the measurement methods.

6.6.3.3 Discussion

An argument for external validity can only partly be crafted by comparing the intra-rater reliability values between different measurement methods. The performance observed during the assessment and the students' answers in the interviews following the assessments turned out to be mutually supportive. Still, a comparison of direct observations with written protocols supported the findings of Gott and Duggan (2002): written protocols and videos seemed to rely on different students' expertise. Important student data found in the written protocols must have been missing to ensure a more compatible rating of both measurement methods.

As Table 6.12 shows, most student pairs improved when solving the CIs in the cognitive lab. All these students had already solved two CIs in school testing. The validated hierarchy among the QSs, assessed by written protocols (samples A–C), was not found in this small sample, where two CIs were solved in one special situation: the eight student pairs reached the fourth QS more frequently and more easily than they had the third QS. One reason may be that these student pairs, since they had already solved two CIs, may have been able to solve the problems more quickly. They likely could have achieved task 4 in time (18 minutes). Furthermore, for students to improve in task 3 (making fair comparisons), special cues and explanations would be required. Students should be told that, according to the rating manual, at least two conditions must be reported or cited to reach the success criterion of QS3.

6.7 Conclusion

Evidence on the structural validity, as well as on the instruments' generalisability have been found in two of the three studies above (RQ1). In addition, we examined wether or not four different measures may improve the instruments' validity (RQ2): working with an a priori progression model (RQ2.1), standardizing the items and the manual (RQ2.2), augmenting the number of tasks per student (RQ2.3), and comparing different assessment and measurement methods (RQ2.4). The first three measures led to an improvement of the instruments' validity. Working with an a priori progression model and introducing three (not four) quality standards helped to relate the scoring system to the construct domain. The same progression of expertise in CIs, which had already partly be validated using item-response theory (Gut et al., 2017), was found here through classical testing. This progression seems to be independent of the content domain.

By standardising as much as possible the item shells and the rating manual and by augmenting the number of tasks to be solved per student, we were able to attain high inter-rater reliability values ($.79 \leq p_0 \leq .98$; $.56 \leq \kappa \leq .97$) and a good instruments' generalisability coefficient ($\rho^2 = 0.72$; $n = 342$). This coefficient is higher than predicted by Miller (1998) and Shavelson and colleagues (1993).

However, none of these three measures helped to reduce the high task-sampling variability that is so often found in PAs, rendering the instrument not practical for evaluating students' expertise in large-scale assessments. In the words of Shavelson, Ruiz-Primo, and Wiley (1999), task-sampling variability remains the Achilles's heel of science-performance assessments.

Furthermore, comparing different assessment and measurement methods (external validity study) showed a weak relationship between test scores from written protocols and direct observations rated with the same manual: students showed lower expertise in written protocols than in direct observations. These two measurement methods must be compatible if the instrument is to be used to evaluate students' expertise in 'doing science'. Evidence on the external validity of the instrument is still missing. Overlooking linguistic skills and students' age may be two reasons that these students did not write down explicitly what they were doing during the investigations. Other reasons may be found when examining the impact of learner variables and the instrument's fairness (i.e. external validity).

Next to the results reported here and used to answer our research questions, a training effect was observed in the study (of external validity) administered in the cognitive lab (sample D).

Student pairs, in general, solved the second CI better than the first one. In most cases, a different hierarchy among the four quality standards was found. The students seemed to improve their performance (without any kind of scaffolding) in some, but not all, of the tasks to be solved. Our next step is to investigate arguments for respectively against our instrument's substantive validity by using direct observations and a larger sample size of students working in a cognitive lab.

Finally, Messicks' idea of eliciting evidence for validity under different aspects (1996), as well as the implementation of different postulated measures, both helped developing and, in particular improved, validating our performance assessment.

7 Beobachten lernen. Aufgaben zur Förderung der Beobachtungskompetenz

Hild, P., Kölbach, E. & Metzger, S. (2015). Beobachten lernen. Aufgaben zur Förderung der Beobachtungskompetenz. *Naturwissenschaften im Unterricht Chemie, 149,* 22–26.

7.1 Einleitung

Naturwissenschaftliche Beobachtungen unterscheiden sich von alltäglichen Sinneswahrnehmungen und Vorgehensweisen vor allem darin, dass sie zielgerichtet (bzw. fokussiert) und kategoriengeleitet[3] sind (Schulz, Wirtz & Starauschek, 2012). Während offen oder unspezifisch formulierte Beobachtungsaufträge (z. B. “Beobachte!”) dazu führen können, dass gerade das wahrgenommen wird, was für die naturwissenschaftliche Beobachtung irrelevant ist, sollen kategoriengeleitete Aufträge den Fokus auf das Relevante richten. Dadurch kann unter anderem sichergestellt werden, dass mehrere Beobachterinnen und Beobachter in derselben Situation unter den gleichen Bedingungen zum selben Resultat gelangen (Pfeifer, 2002). Das Beispiel in Abbildung 7.1 zeigt exemplarisch, was Schülerinnen und Schüler einer 7. Klasse wahrnehmen und notieren, wenn sie einen unspezifischen Beobachtungsauftrag zum Vergleich einer Wasser-in-Öl-Emulsion (Crème A) mit einer Öl-in-Wasser-Emulsion (Crème B) erhalten. Die Vielfalt des Notierten entspricht nicht den Erwartungen der Lehrkraft: Beschreiben die einen eher den Aggregatzustand der beiden Crèmes, befassen sich die anderen eher mit Kategorien wie Farbe, Konsistenz, Geruch oder Glanz. Einige Aussagen enthalten zudem Deutungen, die den Einfluss des Vorwissens bzw. der Alltagserfahrungen deutlich machen. Auffallend ist, dass die Jugendlichen fast nie angeben, von welcher Crème sie gerade sprechen. Dieses Beispiel zeigt eindrücklich, dass Beobachtungen für Lernende nicht trivial sind. “Richtiges” Beobachten muss mit den Schülerinnen und Schülern eingeübt und gezielt gefördert sowie mögliche Schwierigkeiten in der Unterrichtsplanung von Beginn an mit berücksichtigt werden. Außerdem ist zu berücksichtigen, dass das Vorwissen und die Präkonzepte des Beobachters bzw. der Beobachterin einen starken Einfluss auf die Wahrnehmung haben.

[3] Statt von kategoriengeleiteten Beobachtungen wird vor allem in der Biologie häufig von *kriteriengeleiteten Beobachtungen* gesprochen. Allerdings beobachten Schülerinnen und Schüler im Unterricht meistens anhand von Kategorien (wie z. B. Farbe, Temperatur oder pH-Wert) und nicht anhand von Kriterien (wie z. B. “ist höher als” oder “wird erreicht wenn”). Deshalb wird in diesem Beitrag die Bezeichnung kategoriengeleitete Beobachtung verwendet.

Ein Auftrag – viele Beobachtungen

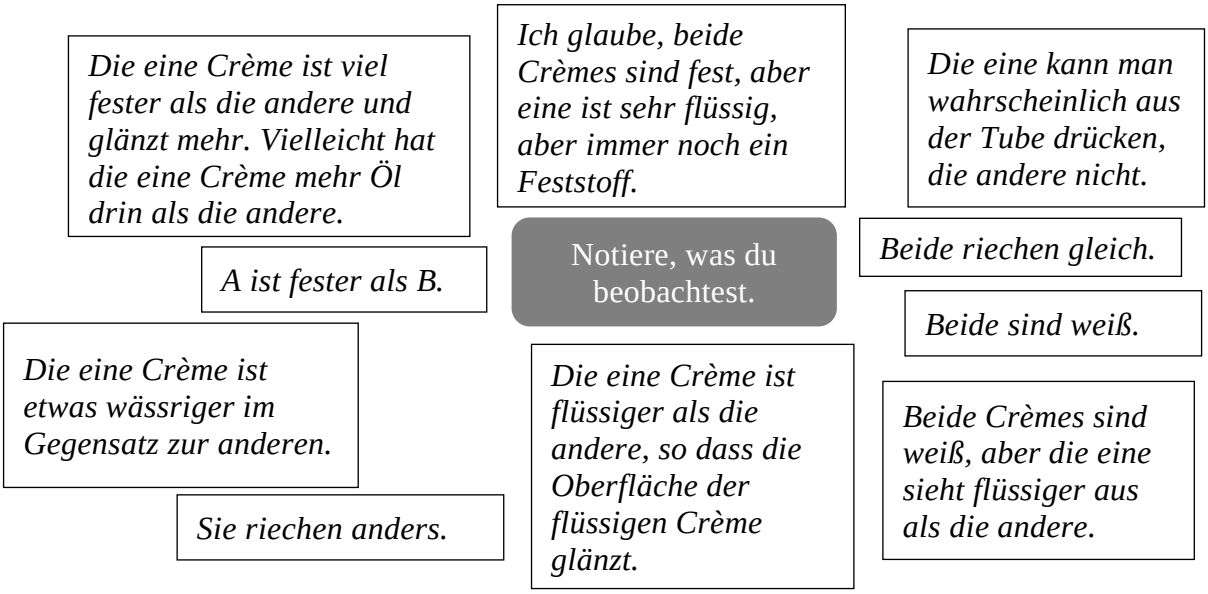


Abb. 7.1 Beispielantworten von Schülerinnen und Schülern zum Auftrag: *Untersuche zwei Crèmes (A und B) ohne sie aus der Plastikschale zu nehmen. Notiere, was du beobachtest.*

7.2 Wie beobachtet man “richtig”?

Das genaue Beobachten von Phänomenen ist eine der grundlegenden Voraussetzungen für die Entwicklung naturwissenschaftlicher Konzepte. Entsprechend sollte die Förderung der Beobachtungskompetenz im naturwissenschaftlichen Unterricht nicht fehlen. In Anlehnung an Arnold et al. (2010) wird Beobachtungskompetenz durch Fähigkeiten und Fertigkeiten beschrieben, mit denen naturwissenschaftliche Fragen durch das Beobachten von Merkmalen, Phänomenen und Prozessen beantwortet werden können. Entsprechend werden unter Beobachtungsaufgaben Aufgaben zur Förderung und Überprüfung kategoriengeleiteter Beobachtungen von Phänomenen und Vorgängen verstanden, die entweder durch Schülerinnen und Schüler oder Lehrkräfte initiiert, arrangiert und manipuliert werden (siehe auch Barzel, Reinhoffer & Schrenk, 2012).

Traditionell ist das Beobachten vor allem in der Biologie verankert: Das Betrachten (statischer Objekte) und das Beobachten (von Prozessen, Zusammenhängen und Wirkgefügen) stellen hier zwei separate Methoden zur Erkenntnisgewinnung dar. Darüber hinaus wird noch zwischen direktem und indirektem Beobachten, mit oder ohne Hilfsmittel unterschieden (z. B. Wellnitz & Mayer, 2008).

Auch wenn im Chemie- und Physikunterricht zumeist andere Formen des Experimentierens wie beispielsweise skalenbasiertes Messen, fragengeleitetes Untersuchen oder effektbasiertes Vergleichen (vgl. Gut et al., 2014) im Vordergrund stehen, sollten Beobachtungsaufträge nicht vernachlässigt werden. Diese stellen zum einen die erste Stufe der naturwissenschaftlichen Erkenntnisweisen (Teilschritt des Experimentierzyklus im Unterricht) dar, können zum anderen aber auch zum Überprüfen von Vermutungen und Hypothesen verwendet werden.

Auch wenn das Beobachten stark kontextabhängig ist, legen erste Ergebnisse der Validierung eines Modells für experimentelle Kompetenzen folgende Kompetenzstufen für das Beobachten nahe (Metzger et al., 2014b): 1. korrektes und vollständiges Beobachten eines Phänomens, 2. Unterschiede zweier Phänomene identifizieren, 3. Gemeinsamkeiten zweier Phänomene identifizieren. Ziel ist es, validierte Kompetenzstufen beschreiben zu können, welche auch die Verwendung eines geeigneten Beobachtungsinstruments einschließen.

7.3 Kompetent beobachten im Chemieunterricht

Im Chemieunterricht können häufig mehrere Beobachtungen gleichzeitig stattfinden. Beispielsweise sollen die Schülerinnen und Schüler beim Beobachtungsauftrag *Königsblau als Indikator* (siehe Aufgabenbeispiele am Ende) eine Säure-Base-Reaktion zwischen Essigsäure und Natriumhydrogencarbonat beobachten. Dabei kann Folgendes simultan wahrgenommen werden: Veränderung des pH-Wertes, des Farbumschlags je nach Indikator, Entstehung von Kohlenstoffdioxid, Veränderung der Temperatur und je nach Menge zugefügter Backhefe eine Fällung. Aus diesem Grund sollte schon bei der Planung einer Beobachtungsaufgabe klar sein, welche Ziele verfolgt und anhand welcher Kategorien die Beobachtungen durchgeführt werden sollen (vgl. Duit, Gropengießer & Stäudel, 2007). Je nach festgelegtem Ziel hängt es stark vom Vorwissen ab, ob und wie die Aufgabe bewältigt werden kann. Beispielsweise kann ohne Vorwissen gesehen werden, dass sich die Farbe einer Essiglösung mit entsprechendem Indikator (hier der Tinte) nach der Zugabe von Backpulver verändert. Jedoch müssen die Schülerinnen und Schüler das Konzept einer Säure-Base-Reaktion verstanden haben, um auf die Idee zu kommen, die Veränderung des pH-Wertes (mithilfe geeigneter Messinstrumente) zu beobachten. Deshalb verlangt das kategoriengeleitete Beobachten – zumindest bei Novizen – Leitfragen oder spezifisch ausformulierte Aufträge wie zum Beispiel: „Was kannst du zur Farbe der Flüssigkeit im Messbecher sagen? Beschreibe und skizziere, was du beobachtest."

Darüber hinaus sind auch bei einfachen Beobachtungsaufträgen Exaktheit, Sorgfalt, Vollständigkeit und Geduld unabdingbar (vgl. Muckenfuß, 1995). Dazu gehört beispielsweise ebenso, dass Schülerinnen und Schüler Beobachtung und Interpretation nicht vermischen und das Beobachtete korrekt protokollieren. Dafür ist es hilfreich, schon im Auftrag darauf hinzuweisen. Außerdem können Lernende mit fehlenden sprachlichen Kompetenzen durch Skizzen oder gegebene Textsegmente unterstützt werden. Bei einigen Beobachtungsaufträgen genügt es nicht, sich auf die menschlichen Sinneswahrnehmungen zu beschränken. Dann werden Hilfsmittel wie beispielsweise Universalindikatorpapier oder ein Mikroskop benötigt, deren Einsatz vorgängig mit den Schülerinnen und Schülern eingeübt werden sollte.

Sollen bewusst mehrere Phänomene oder Vorgänge verglichen werden, muss auf einen fairen Vergleich (gleiche Rahmenbedingungen, Variablenkontrolle) der Beobachtungen geachtet werden. Dafür sollten die Schülerinnen und Schülern schon einige Erfahrung mit dem kategoriengeleiteten Beobachten gesammelt haben.

Methodenkarte: Wie beobachte ich richtig?

Vorbereitung

- ☐ Welches Ziel verfolgst du mit der Beobachtung?
  z. B. ein Phänomen beschreiben, eine Vermutung bestätigen, eine Hypothese überprüfen,…
- ☐ Was weisst du schon über diese Thematik?
  Vielleicht hast du im Unterricht oder außerhalb der Schule schon Ähnliches beobachtet?
- ☐ Was sollst du beobachten?
  z. B. Farbe, Geruch, pH-Wert, Temperaturwechsel,...

Durchführung

- ☐ Arbeite exakt *(beobachte genau und beschreibe präzise)*
- ☐ Arbeite sorgfältig *(z. B. sorgfältiger Umgang mit dem Material)*
- ☐ Notiere alle Beobachtungen, die für das Ziel relevant sind.
  Häufig können mehrere Sachen gleichzeitig beobachtet werden (z. B. Temperaturveränderungen, Farbwechsel, Veränderungen des Aggregatzustands, ...). Wenn du nicht sicher bist, welche Beobachtungen für das Ziel relevant sind, schreibe am besten alle auf.
- ☐ Nimm dir Zeit *(häufig muss man eine Beobachtung wiederholen)*
- ☐ Interpretiere deine Beobachtungen noch nicht
- ☐ Gib an ob du Hilfsmittel verwendet hast *(z. B. Lupe, Mikroskop, pH-Papier, Thermometer,...)*

Reflexion

- ☐ Hast du den Auftrag aus deiner Sicht gut gelöst?
- ☐ Was könntest du zukünftig noch verbessern?

Abb. 7.2 Methodenkarte zum kompetenten Beobachten für Schülerinnen und Schüler.

Auf der Methodenkarte zum Beobachten (Abbildung 7.2) sind Merkmale der Beobachtungskompetenz für den Chemieunterricht zusammengefasst. Je nach Auftrag können diese bei der Planung einer Beobachtungsaufgabe oder der Überprüfung der Beobachtungskompetenz unterschiedlich stark gewichtet werden.

7.4 Aufgabenbeispiele

Im Folgenden werden zwei Aufgaben vorgestellt, mit deren Hilfe das kategoriengeleitete Beobachten eingeübt, vertieft bzw. gefestigt werden kann.

Die Beobachtungsaufgabe *Königsblau als Indikator* (siehe Versuchskarte am Ende des Kapitels 7) bietet die Möglichkeit, Beobachtungen bezüglich mehrerer Kategorien (Temperatur, Farbe, Entstehung eines Gases, evtl. Fällung) zu verfolgen. Somit bietet sie Gelegenheit, die Wichtigkeit des Formulierens von Beobachtungszielen zu thematisieren. Je nach vorgegebenen Zielen werden Beobachtungsinstrumente genutzt und Beobachtungen dokumentiert. Im einfachsten Fall (Beispiel der Versuchskarte) sollen die Lernenden den Farbwechsel beobachten. Gibt man einige Tropfen der pH-abhängigen Tintensorte Königsblau (Pelikan Tinte 4001$^{(R)}$) in warmes (leicht basisches) Leitungswasser, so entfärbt sich die blaue Lösung.

Durch die Zugabe von Essig (Essigsäure) erscheint die Lösung wieder blau. Zuständig für diese Färbung sind die in dieser Tinte enthaltenen Triphenylmethanfarbstoffe, die aufgrund der Struktur des Chromophors Licht im gelben Bereich des Spektrums absorbieren, so dass die Komplementärfarbe blau erscheint. Durch die Zugabe von Backpulver (enthält Natriumhydrogencarbonat) bindet ein Hydroxidion an das zentrale C-Atom des Moleküls. Durch die nun erfolgte sp3-Hybridisierung verkleinert sich das mesomere System und absorbiert Licht im UV-Bereich. Für unser Auge erscheint die Lösung dann farblos[4]. Das Beobachten des Farbwechsels erfolgt somit auf drei Stufen, welche alle von den Schülerinnen und Schülern notiert werden müssen. Zusätzlich wird exaktes Arbeiten verlangt, da die Farbwechsel nur dann deutlich zu beobachten sind. Soll die Schwierigkeit der Aufgabe erhöht werden, so bietet es sich an, mehrere Beobachtungsziele gleichzeitig anzugeben bzw. den Fokus auf Ziele zu richten, die einen Einsatz von zusätzlichen Beobachtungsinstrumenten (hier Thermometer bzw. pH-Papier) erfordern. Ist den Lernenden das Beobachten schon vertraut, so kann die Aufgabe erweitert werden, indem ein Vergleich mit der Tintensorte "Türkis" hinzugenommen wird. Die Tintensorte Türkis (Pelikan Tinte 4001$^{(R)}$) ist kein pH-abhängiger Indikator und verändert weder durch Zugabe von Essig noch von Backpulver ihre Farbe. Bei beiden Tintensorten gibt es jedoch auch Gemeinsamkeiten: Nach Zugabe von

[4] Tintenkiller enthalten eine Flüssigbase, die den gleichen Effekt verursacht und dem "Löschen" der Farbe dient. Andererseits befinden sich in der Tinte des Tintenkillerstiftes keine pH-abhängigen Stoffe, sodass man auch auf basischem Untergrund erneut schreiben kann.

Backpulver schäumen beide Lösungen auf (Säure-Base Reaktion), zudem kann häufig eine Fällung beobachtet werden. Dies liegt daran, dass Backpulver auch Stoffe wie Maisstärke enthält, die im Wasser unlöslich sind und mit keinem anderen Stoff eine Reaktion eingehen.

Mit der Beobachtungsaufgabe *Sind alle Crèmes gleich?* (siehe Versuchskarte am Ende des Hefts) können das "faire Vergleichen" und die Variablenkontrolle eingeübt und gefestigt werden. Diese Aufgabe sollte erst dann genutzt werden, wenn das kategoriengeleitete Beobachten bereits thematisiert wurde. In der Aufgabe sollen zwei Öl-in-Wasser-Emulsionen (Crème A: Körpermilch wie z. B. Balea leichte Bodylotion, Crème C: Handcrème wie z. B. atrix Hand & Nail) mit einer Wasser-in-Öl-Emulsion (Crème B: fettige Gesichtscrème wie z. B. Nivea Crème) verglichen werden.[5] Ziel ist es, Gemeinsamkeiten und Unterschiede der Crèmes zu identifizieren, um eine Aussage treffen zu können, welche der Crèmes sich ähnlicher sind. Zur Zielerreichung gibt es verschiedene Wege. Zum einen hinterlassen wässrige Crèmes einen Wasserrand auf einem Filterpapier, während fetthaltige Crèmes dies nicht oder nur begrenzt tun. Stattdessen hinterlassen diese einen Fettfleck (der tagelang sichtbar bleibt). Weiterhin lösen sich wasserhaltige Crèmes (besser) in Wasser. Als dritte Möglichkeit bietet sich das Einfärben mit Methylenblaukristallen an: Methylenblaukristalle mischen sich mit wässrigen Emulsionen und färben die Crèmes blau. Fettige Crèmes lassen sich nicht mit Methylenblaukristallen mischen, die Kristalle bleiben in den Crèmes deutlich sichtbar. Je nach Niveau kann die Aufgabe so formuliert werden, dass einer bis drei der Vergleiche von den Schülerinnen und Schülern durchgeführt werden müssen. Alternativ (wenn die Lernenden bereits sehr fortgeschritten sind) kann der Auftrag so frei formuliert werden, dass alle Aspekte von den Jugendlichen selbstständig durchgeführt werden sollen. Immer sollte jedoch die Variablenkontrolle und damit der "faire Vergleich" thematisiert werden.

Die beiden vorgestellten Beobachtungsaufgaben können zum Einüben, aber auch als Überprüfungsaufgaben der Beobachtungskompetenz eingesetzt werden. In letzterem Fall kann die Methodenkarte (Abb. 7.2) als Orientierungshilfe für die Schülerinnen und Schüler eingesetzt werden. Anhand des Vorgehens der Lernenden lässt sich dann diagnostizieren, welche Schwierigkeiten beim Beobachten noch auftreten und was noch geübt werden muss. Weitere Aufgaben für den Anfangsunterricht (zum Üben oder zur Überprüfung) sind zum

[5] Dieser Auftrag könnte auch mit Magerquark (Öl-in-Wasser-Emulsion) und Streichkäse (Wasser-in-Öl-Emulsion) durchgeführt werden.

Beispiel der Ölfläschchen-Versuch (Peter, 2007) oder der Aussalzeffekt (Schmidkunz, 1998). Beim Aussalzeffekt befindet sich in einem Reagenzglas eine homogene Lösung aus der gleichen Menge Wasser und Aceton. Durch Zugabe von reichlich Natriumchlorid entsteht ein heterogenes Gemisch (siehe Abbildung 7.3).

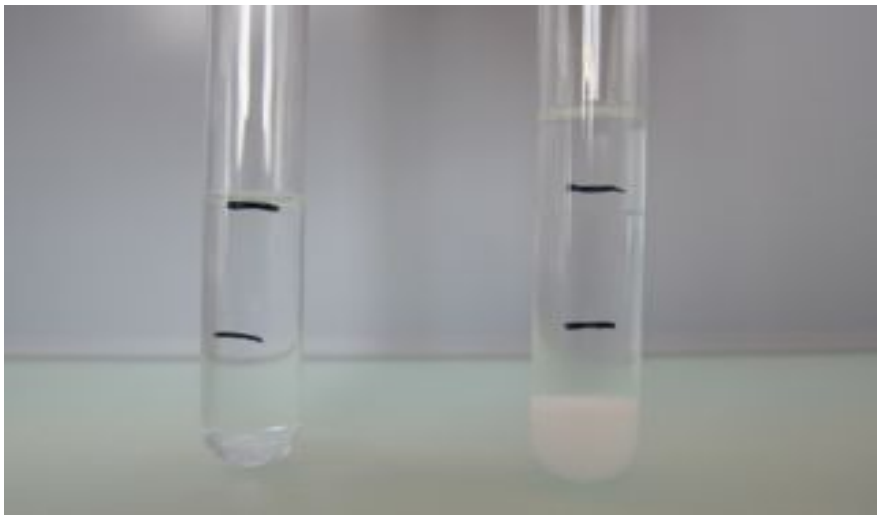


Abb. 7.3 Homogenes (*links*) und heterogenes Gemisch (*rechts*).

Versuch	**Königsblau als Indikator**	**Zeit**
Schüler		**20 Min.**

Geräte

Thermosflasche mit warmem Wasser, 2 Bechergläser (500 ml), Löffel, Holzstab, 2 (Pasteur-)Pipetten mit Sauggummi

Chemikalien

Essig im Schnappdeckelglas, 3 g Backpulver in Tüte, Tinte (Königsblau und Türkis der Marke Pelikan 4001$^{(R)}$)

Durchführung

1) Fülle ein Becherglas zur Hälfte mit warmem Wasser. Füge 2 bis 3 Tropfen der Tintensorte Königsblau hinzu. Rühre, um alles gut zu vermischen. **Beobachte und beschreibe**, was mit der **Tintenfarbe** Königsblau im Becherglas passiert.
2) Füge mit Hilfe der Pipette etwa 30 Tropfen Essig hinzu. Rühre, um alles gut zu vermischen. **Beobachte und beschreibe**, was mit der **Farbe der Flüssigkeit** im Becherglas passiert.
3) Gib ein bis zwei Löffel Backpulver in dasselbe Becherglas. Rühre, um alles gut zu vermischen. **Beschreibe und skizziere**, welchen **Einfluss Backpulver auf die Farbe der Lösung** hat.
4) **Wiederhole** die Schritte 1 bis 3 mit der **Tintensorte Türkis.**
5) **Vergleiche** die beiden Versuche.

Wasser kurz nach Zugabe von Tinte Königsblau

Wasser mit Tinte Königsblau

nach Zugabe von Essig

nach Zugabe von Backpulver

Wasser kurz nach Zugabe von Tinte Türkis

Wasser mit Tinte Türkis

nach Zugabe von Essig

nach Zugabe von Backpulver

Zusammengefasst:

Die Lösung mit Königsblau verändert ihre Farbe nach Zugabe von Essig und nach Zugabe von Backpulver. Die Färbung der Lösung mit Türkis bleibt unverändert.

Beide Lösungen schäumen auf, wenn Backpulver hinzugefügt wird. In beiden Lösungen können, nach Zugabe von Backpulver, feste Stoffe ausfallen.

Abb. 7.4 Versuchskarte *Königsblau als Indikator*.

Versuch	**Sind alle Crèmes gleich?**	**Zeit**
Schüler		**20 Minuten**

Geräte

3 Filterpapiere, Lupe, Plastikspatel, Holzstab, 3 Schnappdeckelgläser, 3 Plastikbecher

Chemikalien

3 Crèmes in Plastikschalen, Wasser in 0,5 l Flasche

Hinweise zu den Crèmes:

A: Körpermilch (z. B. Balea leichte Bodylotion)
B: fettige GesichtsCrème (z. B. Nivea Crème)
C: HandCrème (z. B. atrix Hand & Nail)

Durchführung

1) Fülle drei Bechergläser zur Hälfte mit Wasser und füge zu jedem Glas eine kleine Portion einer Crème hinzu. Rühre. **Beobachte und beschreibe**, wie sich die **3 Crèmes im Wasser** verhalten. *Du kannst den gleichen Auftrag auch in Schnappdeckelgläsern durchführen, die Gläser mit einem Deckel verschließen und dann kräftig schütteln.*
2) Füge je eine kleine Portion der 3 Crèmes auf ein Filterpapier. Warte 1 bis 2 Minuten. **Beobachte und beschreibe**, was **auf dem Filterpapier** passiert.
3) Füge einige Methylenblaukristalle zu den übrig gebliebenen Crèmes in den Plastikschalen. Vermische die Kristalle mit den jeweiligen Crèmes. **Beobachte und beschreibe**, was mit den **Kristallen** passiert.

Crèmes mit Methylenblau

Crèmes mit Wasserrand oder Fettfleck

Crèmes im Wasser nach Rühren

Crèmes im Wasser nach Schütteln

Zusammengefasst:

Crème A und Crème C verhalten sich deutlich ähnlicher als Crème B:

- Methylenblau färbt die Crèmes A und C, hingegen nicht die Crème B (*Bild oben links*)
- Crèmes A und C hinterlassen einen Wasserrand, Crème B hinterlässt einen Fettfleck (*Bild oben rechts*)
- Crèmes A und C trüben das Wasser nach dem Rühren oder nach dem Schütteln, Crème B nicht (*Bilder unten*)

Alle drei Crèmes sind weiß und ähneln sich äußerlich.

Abb. 7.5 Versuchskarte *Sind alle Crèmes gleich?*.

8 Adaptives kompetenzbezogenes Feedback beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten. Eine empirische Untersuchung zur Wirksamkeit unterschiedlicher Feedbackformen

Hild, P., Buff, A., Gut, C. & Parchmann, I. (im Review-Verfahren). Adaptives kompetenzbezogenes Feedback beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten. Eine empirische Untersuchung zur Wirksamkeit unterschiedlicher Feedbackformen. *Zeitschrift für Didaktik der Naturwissenschaften.*

8.1 Zusammenfassung

In der Studie wurde die Wirksamkeit unterschiedlicher adaptiver Feedbackformen auf die Entwicklung experimenteller Kompetenz von Schülerinnen und Schüler aus leistungsschwachen Klassen der Jahrgangsstufe 7 untersucht (n=149, 44.3 % weiblich). Die Beurteilung der Kompetenz und ein daran gekoppeltes adaptives Feedback bezogen sich auf ein validiertes problemtypenbasiertes Kompetenzstufenmodell. Die Schülerinnen und Schüler lösten zu vier Zeitpunkten jeweils eine hands-on Aufgabe. Der erste Zeitpunkt wurde als Prätest, der letzte als Posttest genutzt, während der zweite und dritte Zeitpunkt als Intervention mit unterschiedlichem Feedback wirkte. Schülerinnen und Schüler wurden zufällig einer von drei Interventionsgruppen oder der Kontrollgruppe zugewiesen. Schülerinnen und Schüler aus den drei Interventionsgruppen erhielten zu zwei Zeitpunkten entweder (rückmeldendes) feeding back, (hinweisgebendes) feeding forward oder gleichzeitig beide Formen. Die Ergebnisse zeigen keine statistisch signifikanten Unterschiede in der Entwicklung der experimentellen Kompetenz zwischen Schülerinnen und Schülern der Interventionsgruppe bzw. den einzelnen Interventionsgruppen und der Kontrollgruppe. Das zentrale Problem für die statistisch nicht signifikanten Ergebnisse scheint die zu geringe Größe der Stichprobe zu sein. Insbesondere in einem Falle – Schülerinnen und Schüler mit feeding back – deutet die Effektstärke des Unterschieds zur Kontrollgruppe darauf hin, dass hier Potenzial vorhanden sein könnte, um die experimentelle Kompetenz bei Schülerinnen und Schüler aus leistungsschwachen Klassen der Sekundarstufe I zu fördern. Optimierungsmöglichkeiten mit Blick auf allfällige neue Studien werden diskutiert.

8.2 Scaffolds beim praktisch-naturwissenschaftlichen Arbeiten

Unter praktisch-naturwissenschaftlichem Arbeiten in der Schule werden „…alle beobachtenden und experimentellen Aktivitäten im naturwissenschaftlichen Unterricht…" (Wilhelm & Kunz, 2016, 126) zusammengefasst. Die beiden Autoren Wilhelm und Kunz unterscheiden hier zwischen acht verschiedenen Ausprägungen des praktisch-naturwissenschaftlichen Arbeitens: dem *Betrachten*, dem *Beobachten*, dem *Messen*, dem *Studieren*, dem *Erkunden*, dem *Vergleichen*, dem *Versuchen* und dem *Experimentieren*. Gemäß Wilhelm und Kunz stehen bei diesen Ausprägungen unterschiedliche experimentelle Kompetenzen im Vordergrund. Dem Experimentieren, als hypothesengeleitetes Forschen durch Manipulation der unabhängigen Variable, wird in der Naturwissenschaftsdidaktik einen besonderen Stellenwert zugeschrieben (Emden, Bewersdorff & Baur, 2019). In der vorliegenden Studie lag der Fokus auf einer dieser Ausprägungen − dem effektbasierten Vergleichen (siehe Abschnitt 8.5.2).

Sowohl dem selbstständigen Lernen als auch der individuellen Lernunterstützung werden im Rahmen eines konstruktivistisch orientierten Lehr-Lernverständnisses eine hohe Bedeutung zugeschrieben (Klieme & Warwas, 2011). Unter selbstständigem Lernen wird ein zunehmend selbst regulierter Aufbau und Transfer von Kompetenz(en) verstanden (Friedrich & Mandl 1992). In wieweit Lernunterstützungen Kompetenzen beim selbstständigen Lernen bestärken hängt von unterschiedlichen Faktoren ab (Grad der Autonomie, Art der Lernunterstützung, sowie Kompetenzumfang − siehe Details bei Vorholzer & von Aufschnaiter, 2019). Die Frage, wie (und auch mit welchen Hilfsmitteln) Lehrpersonen individuelle Lernunterstützungen zur Förderung von selbstständigem Lernen planen und einsetzen sollen wurde schon früher aufgeworfen, aber bis heute nicht ausreichend Evidenz basiert beantwortet (Krammer, 2009). Die vorliegende Studie setzt hier an, untersucht unterschiedliche Lernunterstützungen beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten mit dem Ziel Lösungsansätze für den Unterricht zu liefern. Gerade beim praktisch-naturwissenschaftlichen Arbeiten sind, wie bereits oben angedeutet, die geforderten Kompetenzen sehr vielfältig. Dazu zählen fachspezifisch inhaltliche ebenso wie überfachlich methodische Fertigkeiten und Fähigkeiten (Hild et al., 2018c). Durch unterschiedliche *Scaffolds* können diese Kompetenzen für einzelne Schülerinnen oder Schüler gezielt gefördert werden, um ein zunehmend selbstständigeres Anwenden zu ermöglichen.

Gemäß Pea (2004) und Stone (1998) werden (individuelle) Lernunterstützungen als Scaffolds bezeichnet, wenn ein interaktiver Prozess zwischen Lehrperson und Schülerinnen und Schülern stattfindet. Scaffolds sollten immer adaptiv, sprich an den Lernstand und Lernfortschritt der Schülerinnen und Schüler angepasst sein (Glaser, 1972). Diese Adaptivität ist im Unterricht jedoch nicht immer und nicht für alle einzelnen Schülerinnen und Schüler gewährleistet, da die Lehrperson den Lernstand und die Leistungen nicht immer einschätzen und beurteilen kann. So wird in der Literatur zwischen *hard* (statischen, die Schülerprobleme antizipierenden) und *soft* (dynamischen, situativen) Scaffolds unterschieden (siehe Details bei Arnold et al., 2017).

In der naturwissenschaftsdidaktischen Forschung konnte gezeigt werden, dass Scaffolds das selbstständige Lernen im naturwissenschaftlichen Unterricht unterstützen (z. B. Azevedo et al., 2005; Chiu, Chou & Liu, 2002; Mercer et al., 2004; Reigosa & Jimenez-Aleixandre, 2007; Wu & Krajcik, 2006). Gemäß van de Pol können sechs Maßnahmen zur Lernunterstützung auch als Scaffolds eingesetzt werden: *feeding back*, *hints*, *instructing*, *explaining*, *modeling* und *questioning* (siehe Abb.8.1 und Details bei van de Pol et al., 2010). *Lösungshinweise* (hints) zum Beispiel, können bereits in den Aufgaben integriert sein (vgl. Chen et al., 2016; Koenen et al., 2017). Andere Scaffolds hingegen, wie z. B. der Einsatz *gestufter Lernhilfen*, werden den Schülerinnen und Schülern erst während eines Lernprozesses bei Bedarf abgegeben (Wodzinski & Stäudel, 2009). Abbildung 8.1 zeigt eine Zusammenfassung der sechs Scaffolding-Maßnahmen die auch beim praktisch-naturwissenschaftlichen Arbeiten eingesetzt werden können. Die Abbildung zeigt eine exemplarische Auswahl von Interventionsstudien, die den Einfluss von einer oder mehreren Scaffolding-Maßnahmen auf die Entwicklung experimenteller Kompetenzen von Schülerinnen und Schülern bzw. Studierenden untersuchten. Bei den erwähnten Studien werden Lernende vor allem beim Experimentieren, oder bei Teilschritten des Experimentierens unterstützt.

<table>
<tr><th>questioning
Fragen stellen</th><th>explaining
Ziele klären</th><th>feeding back
rückmelden</th><th>modeling
vorzeigen</th><th>hints
Hinweise geben</th><th>instructing
auffordern</th></tr>
<tr><td>Experimentieren
Schreiber & Theyßen 2019

Daten auswerten
Coelho & Séré 1998</td><td>Daten auswerten
Lippmann-Kung 2005; Priemer et al. 2019; Volkwyn et al. 2008</td><td>Experimente planen
Wollenschläger 2012</td><td>Experimentieren
Koenen 2014</td><td>Experimentieren
Wodszinski & Stäudel 2009</td><td>Variablenkontroll-strategie
Kuhn & Dean 2005; Schwichow 2015</td></tr>
<tr><td colspan="2"></td><td colspan="4">Experimentieren
Arnold 2015; Habig et al. 2018; Marschner, 2011; Wahser 2008; Walpuski 2007</td></tr>
<tr><td colspan="6">Experimente planen
Scheuermann 2017</td></tr>
<tr><td colspan="2">feeding up</td><td>feeding back</td><td colspan="3">feeding forward</td></tr>
</table>

Abb. 8.1 Unterschiedliche Scaffolding-Maßnahmen nach van de Pol et al. (2010)

Feeding back wird gemäß van de Pol et al. (2010) zu den Scaffolding-Maßnahmen hinzugezählt und kann als soft Scaffold eingestuft werden, da Schülerinnen und Schüler Informationen zu ihrem individuellen und aktuellen Lernstand erhalten (dynamisch, situativ). Feeding back kann beispielsweise darin bestehen anzugeben, ob eine Aufgabe richtig oder falsch gelöst wurde, wobei damit auch gleich die richtige Antwort mitgeliefert werden kann (z. B. Jaehnig & Miller, 2007). In der Feedbackliteratur wird üblicherweise zwischen feeding back (how am I going?), feeding forward (where to next?) und feeding up (where am I going?) unterschieden (siehe Details bei Hattie & Timperley, 2007; Hattie & Wollenschläger, 2014). Durch das Beachten dieser drei Formen werden Schülerinnen und Schüler dementsprechend über den Lernstand, über das Lernziel und über nächste Schritte informiert (Hattie, 2008). Abbildung 8.1 zeigt, dass sich der Scaffolding Ansatz von van de Pol gut in den Feedback-Ansatz von Hattie & Timperley eingliedern lässt, wobei die beiden Maβnahmen *questioning* und *explaining* zum *feeding up* und die drei Maβnahmen *modeling, hints* und *instructing* zum *feeding forward* hinzugezählt werden können (van de Pol et al., 2010 vs. Hattie & Timperley, 2007). Die Wirksamkeit von Feedback (allgemein) auf den Lernerfolg wurde u. a. von Hattie (2008) in seiner groß angelegten und häufig zitierten Metaanalyse im Allgemeinen belegt und soll im Abschnitt 8.4 für den Bereich experimenteller Kompetenzen im Besonderen ausführlicher beschrieben werden. In der vorliegenden Studie wurde die Wirkung von kompetenzbezogenem (resp. kompetenziellem) Feedback (feeding back vs. feeding forward) beim praktisch-naturwissenschaftlichen Arbeiten untersucht.

8.3 Kompetenzstufenmodellen als wichtige Lern- und Lehrunterstützungen

Kompetenzen sind in Anlehnung an Friedrich & Mandl (1992) stets in konkreten Inhalten eingebettet und können über die drei Stufen *Automatisation*, *Flexibilisierung* und *Transfer auf neue Aufgabenbereiche* selbstständig erworben werden. Sie können sich dabei grundsätzlich nach mehreren Progressionslogiken weiterentwickeln (Adamina & Hild, 2019). Schülerinnen und Schüler werden kompetenter, indem sie lernen, komplexere Probleme zu lösen, bestimmte Probleme qualitativ besser, eigenständiger (Automatisation und Flexibilisierung), in mehr fachlichen Kontexten (Transfer) oder stabiler zu lösen (Gut et al., 2014).

Kompetenzmodelle, welche primär als Messmodelle zu Testzwecken entwickelt wurden, sind laut Reusser (2015) wichtige Lern- und Lehrunterstützungen beim Erwerb und beim Aufbau von Kompetenzen. So können die Ausformulierungen der zu erreichenden Kompetenzen als Lernzielvorgaben und die Modelle selbst, beispielsweise für die Planung und Diagnose von Lernaufgaben genutzt werden. Nützlich sollten sich vor allem Kompetenzstufen- und Kompetenzentwicklungsmodelle erweisen. Diese gehen von gestuften Kompetenzen aus und zeigen auf, unter welchen Bedingungen eine bestimmte Stufe erreicht wird (Schecker & Parchmann, 2006). Validierte Kompetenzstufen- oder Kompetenzentwicklungsmodelle ermöglichen somit, den Schülerinnen und Schülern adaptive, auf die in Testaufgaben geforderten Kompetenzen bezogene Lernunterstützungen zu erteilen (Harks et al., 2014; Wollenschläger et al., 2012).

Im Bereich des praktisch-naturwissenschaftlichen Arbeitens (vor allem für das Experimentieren) wurden solche Modelle postuliert und vielfach empirisch validiert (vgl. Hammann et al., 2007; Mayer et al., 2007; Schreiber, Theyβen & Schecker, 2009). Im Modell des Projekts *Experimentelle Kompetenzen in den Naturwissenschaften (ExKoNawi)* (z. B. Gut et al., 2014) wurden für unterschiedliche experimentelle Problemtypen, theoretisch abgeleitete und hierarchisch angeordnete Kompetenzstufen *a priori* formuliert (Gut et al., 2017). Diese Stufen bezogen sich vor allem auf die nötigen fachmethodischen Kompetenzen (u. a. Variablenkontrollstrategie oder Strategie der Messwiederholung), welche Schülerinnen und Schüler brauchen, um gewisse experimentelle Probleme lösen zu können. Die Ergebnisse der Validierungsstudien (siehe Abschnitt 8.5) zeigten deutlich, dass die Mehrzahl der Schülerinnen und Schüler aus tiefen Jahrgängen und Klassen mit tiefen Anforderungsniveaus (Haupt-/Realschule) in allen gemessenen Experimentiersituationen einen erheblichen Lernbedarf resp. Lernunterstützungen nötig haben (Hild et al., 2018a; 2018c). Scaffolds,

welche sich entlang der Stufen des zur Diagnose verwendeten Modells richten, könnten den Erwerb und Aufbau fehlender Kompetenzen besonders unterstützen.

8.4 Adaptives kompetenzbezogenes Feedback

8.4.1 Praktisch-naturwissenschaftliches Arbeiten ohne Lernunterstützung

In der Vergangenheit wurde deutlich, dass Schülerinnen und Schüler beim praktisch-naturwissenschaftlichen Arbeiten ohne jegliche Art von Lernunterstützung (z. B. beim mehrmaligen Lösen ähnlicher hands-on Aufgaben in large-scale assessments) ihre Kompetenzen nicht verbessern konnten (vgl. Germann & Aram, 1996; Ramseier et al., 2011; Shavelson, 1992). Es fand also kein Kompetenzzuwachs statt. Nebst der fehlenden Lernunterstützung werden häufig auch die starke Kontextabhängigkeit der einzelnen Aufgaben, sowie eine zu wenig hohe Standardisierung der Testhefte wie auch Kodiermanuale als Hauptgründe für den nicht beobachtbaren Zuwachs postuliert (vgl. Gao et al., 1994; Gut, 2012; Hild et al., 2018b; Stecher et al., 2000). Die validierten Testaufgaben aus dem Projekt ExKoNawi bildeten hier eine Ausnahme – eine günstige Kompetenzentwicklung durch mehrmaliges Lösen ähnlicher hands-on Aufgaben war hier bei der Mehrzahl der Schülerinnen und Schüler feststellbar (Hild et al., 2018a).

8.4.2 Praktisch-naturwissenschaftliches Arbeiten mit Lernunterstützung

Beim praktisch-naturwissenschaftlichen Arbeiten mit Lernunterstützung hingegen, konnten signifikante Unterschiede (von mittlerer bis großer Effektstärke) – im Vergleich zu einer Kontrollgruppe ohne Lernunterstützung – bei der Kompetenzentwicklung festgestellt werden: So wurde gezeigt, dass sich gewisse Unterstützungen positiv auf die *Variablenkontrollstrategie*, das *selbstregulierte Lernen*, das *Planen von Experimenten* sowie das *wissenschaftliche Denken* auswirkten (vgl. Arnold et al., 2017; Marschner, 2011; Sandoval & Reiser, 2004; Schwichow et al. 2016; Wollenschläger et al., 2012). Ähnliche Ergebnisse werden auch beim praktisch-naturwissenschaftlichen Arbeiten mit individuellem adaptivem Feedback erwartet und konnten auch gefunden werden (z. B. Scheuermann, 2017).

8.4.3 Praktisch-naturwissenschaftliches Arbeiten mit feeding back und feeding forward

Allgemein zählt Feedback bei „korrektem" Einsatz mit zu den wirkungsvollsten Interventionen zur Förderung von Lern- und Entwicklungsprozessen (Black & Wiliam, 1998; Müller & Ditton, 2014). Kingston & Nash (2011) fanden in ihrer groß angelegten Metaanalyse mit über 300 Studien, dass die Form des Feedbacks einen erheblichen Einfluss auf die Leistungssteigerung hatte (siehe weitere Hinweise hierfür bei Bangert-Drowns et al.,

1991; Kluger & DeNisi, 1996; Marschner, 2011 oder Rakoczy et al., 2008). So konnten etwa auch Kulhavy et al. (1990) zeigen, dass rückmeldende Angebote (feeding back) wirksamer für den Leistungszuwachs von jungen Schülerinnen und Schülern waren als hinweisgebende Angebote (feeding forward). Ihre Vermutung für dieses Ergebnis war, dass scheinbar viele Schülerinnen und Schüler Angebote vor allem nutzen um herauszufinden, ob sie in einer Testaufgabe richtig lagen oder nicht und sich kaum mit dem Ausarbeiten alternativer Lern- bzw. Lösungswege beschäftigen wollten. Diese Vermutung wurde durch weitere Studien gestützt, welche nur mit feeding forward intervenierten und keinen Einfluss auf eine Leistungssteigerung bei Schülerinnen und Schüler feststellen konnten (Harks et al., 2014; Perfetto et al., 1983). Dem feeding forward wird häufig vorgeworfen, es mangele ihm an Adaptivität, sprich, das Angebot könne nicht an den Lernstand und -fortschritt einzelner Schülerinnen und Schüler angepasst sein.

Schon in den Achtziger-Jahren des letzten Jahrhunderts berichtete Bloom (1984) Effektstärken für Interventionsstudien mit feeding back (FB) und feeding forward (FF). Basierend auf den Daten seiner Metastudie postulierte er, dass sich die Effektstärken von FB und FF zueinander additiv verhalten sollten und Interventionen mit gleichzeitigem Einsatz von FB und FF erfolgsversprechender seien als Interventionen mit FB oder FF. Bezogen auf das selbstständige Experimentieren, konnte Scheuermann (2017) diese These in ihrer Studie belegen: Eine Intervention mit gleichzeitigem FB und FF führte hier zum größeren Kompetenzzuwachs bzgl. *naturwissenschaftlichen Arbeitsweisen*. Ein Aspekt, der jedoch gegen die Additivität von FB und FF spricht, ist die Textmenge des Feedback-Angebots. Durch eine Kombination von unterschiedlichen Feedbackformen wird die Informationsmenge deutlich vergrößert, was auch zu einer Überforderung im Sinne einer zu großen kognitiven Belastung bei schwächeren Schülerinnen und Schüler führen könnte (vgl. Atkinson et al., 2000; Kulhavy & Stock, 1989).

Auch die folgende Studie untersuchte die Wirkung von adaptivem kompetenzbezogenen Feedback (AKF) beim praktisch-naturwissenschaftlichen Arbeiten. Der Fokus lag vor dem Hintergrund des bisherigen Erkenntnisstands darauf, Schülerinnen und Schüler aus leistungsschwachen Klassen beim praktisch-naturwissenschaftlichen Arbeiten mit Aufgaben zum effektbasierten Vergleichen durch AKF zu unterstützen. Dazu wurden in einer Intervention drei Experimentalgruppen gebildet: Schülerinnen und Schüler erhielten entweder ein FB, ein FF oder eine Kombination aus FB und FF (siehe Abschnitt 8.5).

8.4.4 Umfang und Intensität von Scaffolding-Maßnahmen

Viele Interventionsstudien, die Scaffolding-Maßnahmen und ihre Wirksamkeit auf die Kompetenzentwicklung beim praktisch-naturwissenschaftlichen Arbeiten untersuchen, sind langfristig angelegt (≥ 8 Lerneinheiten, vgl. Arnold et al., 2017; Chiu, Chou & Liu, 2002; Reigosa & Jimenez-Aleixandre, 2007; Sandoval & Reiser, 2004; Tabak & Baumgartner, 2004; Wu & Krajcik, 2006). Bei diesen Studien sind die verwendeten Kontexte häufig selber Teil des Lerninhalts (z. B. Experimentieren bezogen auf Säuren und Basen bei Walpuski & Sumfleth, 2007; Experimentieren bezogen auf Enzymatik bei Arnold et al., 2017, Experimentieren bezogen auf Metalle und ihre Eigenschaften bei Scheuermann, 2017), mehrere Scaffolding-Maßnahmen werden miteinander verglichen (siehe Abb. 1, z. B. Arnold et al., 2017; Marschner, 2011; Scheuermann, 2017 oder Walpuski & Sumfleth, 2007) und nebst fachspezifisch inhaltlichen Kompetenzen werden weitere Lernervariablen (Selbstwahrnehmung, Kompetenzerleben, aktuelle Lernmotivation, Akzeptanz des Angebots) mit erhoben. Studien, die sich auf Teilkompetenzen (wie hier methodische Kompetenzen beim effektbasierten Vergleichen) beschränken, intervenieren jedoch auch mit weniger als 2 Lerneinheiten (0.6h bei Azevedo et al., 2005; 1.8h bei Chi et al., 2001; 0.5h bei Schwichow et al., 2016; 1.5h bei Marschner, 2011; 1h bei Wahser & Sumfleth, 2008; 1.5h bei Wollenschläger et al., 2012). Des Weiteren variieren bei allen erwähnten Studien die gewählte Stichprobengröße (z. B. n = 50 bei Wollenschläger, 2012 vs. n = 234 bei Scheuermann, 2017), sowie das Anforderungsniveau (z. B. gymnasiale Schülerinnen und Schüler der 11ten Klasse bei Arnold, 2015 vs. Schülerinnen und Schülern aus Real- und Hauptschulen sowie Gymnasien der Klasse 8 bei Schwichow, 2015) stark. Ausser der Interventionsstudie von Wodszinski & Stäudel (2009) sind den Autorinnen und Autoren keine Studien (bezogen auf das praktisch-naturwissenschaftliche Arbeiten) bekannt, welche ausschliesslich bei Schülerinnen und Schülern aus Real- und Hauptschulen intervenierten. Bei vielen der erwähnten Studien werden zudem keine Effektstärken berichtet. Und letztlich fehlen auch Studien, die die Wirksamkeit von Scaffolding-Maßnahmen bei anderen Modi der Erkenntnisgewinnung (wie z. B. hier dem effektbasierten Vergleichen) des praktisch-naturwissenschaftlichen Arbeitens untersuchen.

8.5 Studie zur Wirksamkeit von adaptivem kompetenzbezogenen Feedback

8.5.1 Fragestellungen der Pilotstudie

Die bisherigen Studien zu AKF unterscheiden sich hinsichtlich Stichprobengröβe, Anforderungsniveau, Interventionsdauer, Anzahl Scaffolding-Maßnahmen erheblich (siehe 8.4.4). Ebenso ist die Befundlage keineswegs einheitlich (siehe 8.4.1-8.4.3). Aufbauend auf theoretischen Überlegungen und gestützt auf dem empirischen Forschungsstand sichere (gerichtete) Hypothesen zu formulieren, ist daher kaum möglich. Vor diesem Hintergrund wird in der vorliegenden Studie darauf verzichtet und es werden lediglich vier Fragestellungen formuliert. Entsprechend hat die Studie primär einen explorativen Charakter.

- *Fragestellung 1. Überprüfung der Kompetenzentwicklung ohne Lernunterstützung.* Zeigen Schülerinnen und Schüler auch ohne Lernunterstützung eine Veränderung in der Entwicklung ihrer experimentellen Kompetenz (z. B. Hild et al., 2018a).
- *Fragestellung 2. Feedback vs. kein Feedback.* Unterscheiden sich Schülerinnen und Schüler ohne Lernunterstützung in der Entwicklung ihrer experimentellen Kompetenz von (bzw. der Gesamtgruppe der) Schülerinnen und Schüler mit Lernunterstützung? (z. B. Wollenschläger et al., 2012)?
- *Fragestellung 3. Einfluss der Feedbackform.* Unterscheiden sich Schülerinnen und Schüler aus unterschiedlichen Interventionsgruppen (FB, FF, FB + FF) hinsichtlich ihrer Entwicklung der experimentellen Kompetenz je einzeln von den Schülerinnen und Schüler ohne Lernunterstützung (z. B. Scheuermann, 2017)?
- *Fragestellung 4. Feeding back und feeding forward vs. feeding back.* Unterscheiden sich Schülerinnen und Schüler mit FB + FF in ihrer Entwicklung experimenteller Kompetenz von Schülerinnen und Schüler mit FB (vgl. Bloom, 1984 vs. Kulhavy & Stock, 1989)?

8.5.2. Ausgangslage

In der vorliegenden Studie wurde die Wirksamkeit von unterschiedlichem AKF auf die Entwicklung experimenteller Kompetenz von Schülerinnen und Schüler der Sekundarstufe I untersucht. Schülerinnen und Schüler lösten zu 4 Messzeitpunkten (t_1 bis t_4) jeweils eine hands-on Aufgabe. Die Entwicklung experimenteller Kompetenz bezog sich hier auf den Prä/Post Vergleich (t_1/t_4). Unter experimenteller Kompetenz wurde hier wie im Projekt ExKoNawi das Vorhandensein einer Vielzahl an (vor allem fachmethodischen) Fähigkeiten

und Fertigkeiten verstanden, welche je nach experimentellem Problemtyp (kategoriengeleitetes Beobachten, skalenbasiertes Messen, effektbasiertes Vergleichen, fragengeleitetes Untersuchen) unterschiedlich wichtig für die Problemlösung sind (für Details siehe Gut & Mayer, 2018). Diese Studie befasste sich exemplarisch mit dem experimentellen Problemtyp effektbasiertes Vergleichen (siehe Abb. 8.2).

Die Schülerinnen und Schüler können …

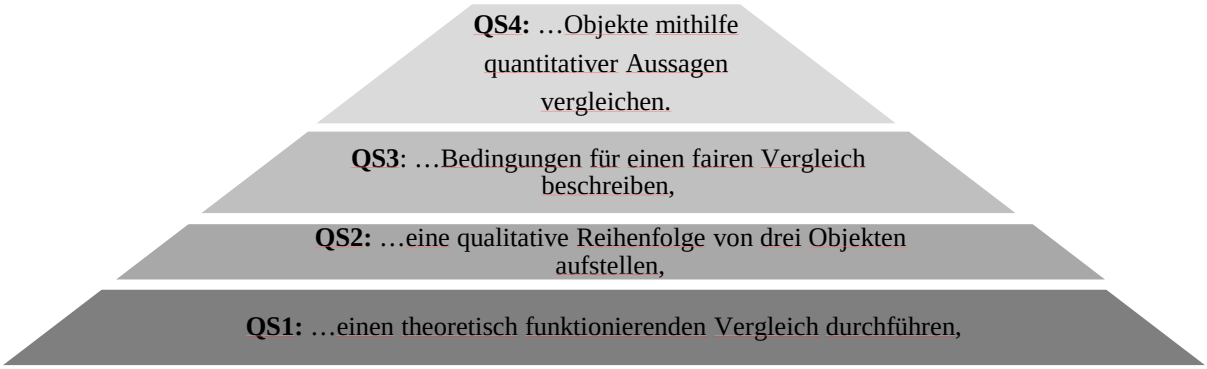


Abb. 8.2 Die vier Qualitätsstandards beim *effektbasierten Vergleichen* – ein Kompetenzstufenmodell.

Aufgaben zum experimentellen Problemtyp effektbasiertes Vergleichen werden definiert als Aufgaben, bei denen Objekte anhand einer gegebenen Eigenschaft verglichen werden sollen und alle anderen Variablen konstant gehalten werden (Solano-Flores & Shavelson, 1997). Im Gegensatz zum kriteriengeleiteten Vergleichen (z. B. Wellnitz & Mayer, 2008), liegt der Fokus bei diesem Problemtyp stärker auf der Erzeugung von Effekten und weniger stark auf dem Erfassen von Gemeinsamkeiten und Unterschieden bezüglich ausgewählter Kriterien. Beim kriteriengeleiteten wie auch beim effektbasierten Vergleichen müssen Rahmenbedingungen gleich gehalten werden und die Vergleiche müssen *fair* sein. Für diesen Problemtyp wurden vier Kompetenzstufen, die wir hier im Weiteren *Qualitätsstandards* (QS) nennen, *a priori* festgelegt (Abb. 8.2). Das eingesetzte AKF in der hier vorgestellten Studie richtete sich entlang dieser QSs: Beim ersten und zweiten Qualitätsstandard wurde überprüft, ob der vorgeschlagene Vergleich sich theoretisch eignet (für zwei bzw. drei Objekte), um eine gewisse Objekteigenschaft zu überprüfen. Es mussten jedoch keine Daten vorhanden sein und der Vergleich durfte auch ohne Variablenkontrolle durchgeführt werden. QS 3 wurde nur dann erreicht, wenn beim Vergleich explizit erkenntlich wurde, dass gewisse Bedingungen eingehalten wurden (Variablenkontrolle). Der letzte QS konnte nur erreicht werden, wenn

quantitative Daten vorhanden waren, die Aussagen über die Ausprägung oder Differenzen der Ausprägungen der Objekteigenschaften zuließen. Die abgebildete Kompetenzstufung entlang der vier QSs konnte bei insgesamt acht unterschiedlichen Aufgaben zu diesem Problemtyp sowie bei allen Jahrgängen und Anforderungsniveaus der Sekundarstufe I gemessen und validiert werden (z. B. Gut et al., 2017, Hild et al., 2018a).

Auch in der vorliegenden Studie wurde diese Kompetenzstufung (das gleiche Muster entlang der QS) vorgefunden: Abbildung 3 zeigt die Progressionen entlang der Qualitätsstandards nach dem Prätest bei den vier hands-on Aufgaben aus der hier vorliegenden Studie (siehe Details zu den Kontexten in Tab. 8.1 und zum Design in Abschnitt 8.6.2). Schülerinnen und Schüler erreichen in allen hands-on Aufgaben häufiger den ersten als den zweiten und häufiger den zweiten als den dritten QS. Der vierte QS wurde im Prätest fast nie erreicht. Die beiden ersten Qualitätsstandards (QS) unterscheiden sich in allen Fällen signifikant (Wilcoxon Test). Die beiden letzten QS wurden fast nie erreicht und sind dementsprechend auch nicht trennscharf.

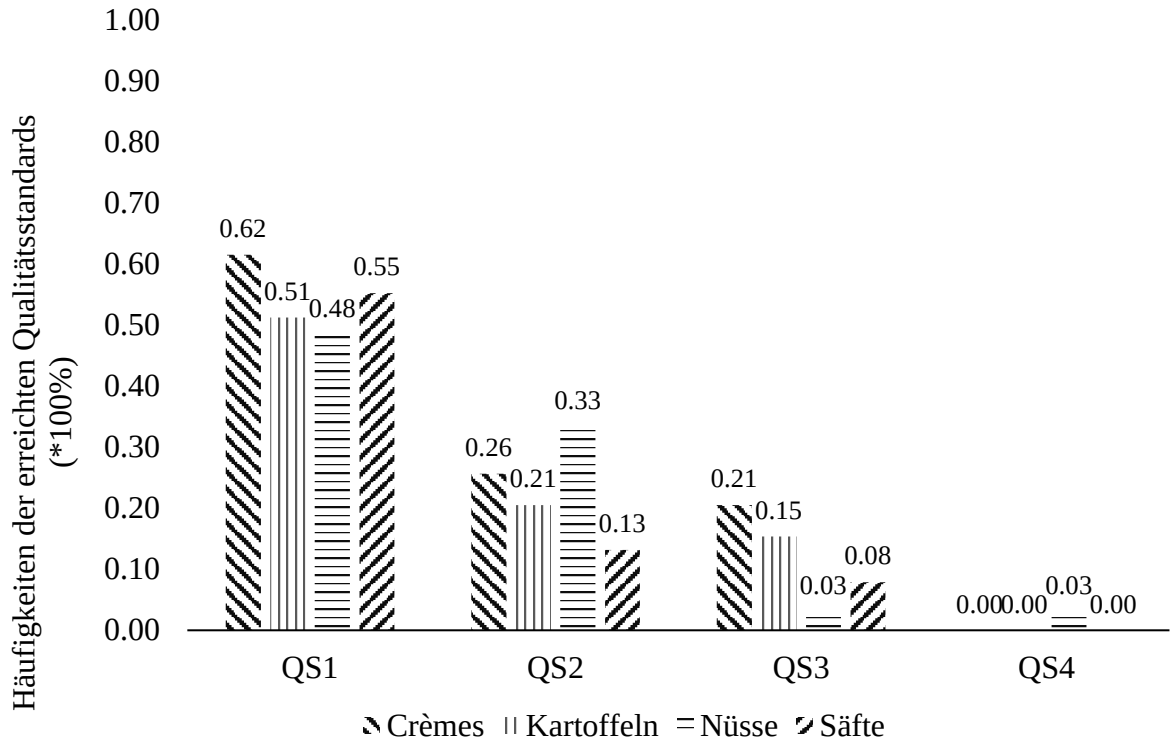


Abb. 8.3 Häufigkeiten der erreichten QSs nach Prätest (t_1, n = 149)

8.6 Methode

8.6.1 Stichprobe

Die Stichprobe bestand aus 185 zwölf- bis fünfzehnjähriger Schülerinnen und Schüler der 7. Klassen aus der Agglomeration der Stadt Zürich (Alter: M = 13.4, SD = 0.55, 44.3 % weiblich). In Zürich gibt es insgesamt 4 unterschiedliche Anforderungsniveaus (Gymnasium, Sek A, Sek B und Sek C). Die Stichprobe bestand ausschließlich aus Schülerinnen und Schülern aus den beiden tiefsten Niveaus (B und C). Alle Schülerinnen und Schüler erhielten vor dem Lösen der ersten Aufgabe einen demographischen Fragebogen (Alter, Geschlecht, Sprache(n)). Als Erstsprache wurden insgesamt 29 unterschiedliche Sprachen angegeben, wobei Deutsch/Schweizerdeutsch als eine Sprache behandelt wurde. 47 % aller Schülerinnen und Schüler gaben an, zuhause mehrsprachig aufzuwachsen. Bei 49.7 % aller Schülerinnen und Schüler wurde zuhause auch Deutsch/Schweizerdeutsch gesprochen, wovon insgesamt 19 Schülerinnen und Schüler angaben, zuhause nur Deutsch/Schweizerdeutsch zu sprechen. Die Schülerinnen und Schüler besuchten 12 Klassen, geführt von 11 Lehrpersonen aus 4 unterschiedlichen Schulhäusern. Für die Überprüfung der Fragestellungen wurde mit einer reduzierten Stichprobe aus 149 (von 185) Schülerinnen und Schüler gearbeitet: 36 Schülerinnen und Schüler (19.4 %) mussten aus der Studie ausgeschlossen werden, 3 Schülerinnen und Schüler wegen mangelnder Deutschkenntnisse und 33 weitere, weil sie an einem oder mehreren Messzeitpunkten fehlten. Da die Intervention insgesamt kurz war (2. und 3. Messzeitpunkt), mussten Schülerinnen und Schüler am ersten und zweiten Messzeitpunkt anwesend sein, um überhaupt adaptive Rückmeldungen zu erhalten. Aus diesem Grund hätten von den 33 aufgrund von Abwesenheit(en) ausgeschlossenen Schülerinnen und Schüler, 18 keine Rückmeldungen erhalten und 15 lediglich eine. Bei einer so kurzen Intervention auch Schülerinnen und Schüler mit lediglich einer adaptiven Rückmeldung einzubeziehen, erschien uns sachlich nicht gerechtfertigt. Da es sich um einen groβen Dropout handelt, wurde im Nachhinein überprüft, ob sich die Teilnehmerinnen und Teilnehmern an der Studie (reduzierte Stichprobe) von den ausgeschlossenen Schülerinnen und Schülern hinsichtlich einiger, für alle zur Verfügung stehender Variablen, unterschieden (siehe 8.6.4).

8.6.2 Design und Treatment

Zu vier Zeitpunkten (t_1 bis t_4) lösten die Schülerinnen und Schüler in Einzelarbeit und jeweils zufällig eine von vier hands-on Aufgaben zum effektbasierten Vergleichen (Tab. 8.1). Für die Bearbeitung der Aufgaben und Dokumentation der Problemlösung hatten sie jeweils achtzehn Minuten Zeit. Danach wurden die Testbögen resp. Schülerprotokolle eingesammelt und ausgewertet. Es wurde darauf geachtet, dass zu jedem Zeitpunkt etwa gleich viele Schülerinnen und Schüler pro Klasse die gleiche Aufgabe lösten (siehe Details unter 8.6.3). Zwischen den jeweiligen Zeitpunkten lagen 1 bis 3 Tage. Zur Beantwortung der Fragestellungen wurden die Schülerprotokolle vom Zeitpunkt t_1 als Prätest und diejenigen vom Zeitpunkt t_4 als Posttest verwendet.

Tab. 8.1 Feeding up (Zielklärung) zu den vier hands-on Aufgaben aus der Studie.

Testaufgabe	**Feeding up**
Kartoffeln	*Isabelle und Michael wollen herausfinden, welcher Stoff (Salz, Zucker, Mehl) mehr Wasser aus einer Kartoffel ziehen kann.*
Crèmes	*Otto und Olivia wollen herausfinden, welche Crème (Körperlotion, Handcrème, Nivea®) am wässrigsten ist.*
Säfte	*Mia und Nino wollen herausfinden, welche Frucht (Gurke, Apfel, Zucchini) am meisten Saft enthält.*
Nüsse	*Xenia und Xaver wollen herausfinden, welche Nuss (Walnuss, Erdnuss, Mandel) am meisten Fett enthält.*

Die Intervention fand während der Zeitpunkte t_2 und t_3 statt. Alle Schülerinnen und Schüler wurden für beide Treffen einer von drei Interventionsgruppen (IG1, IG2 und IG3) oder der Kontrollgruppe (KG) zugeteilt und erhielten jeweils vor dem Lösen einer Testaufgabe eine Infokarte, welche sie vor dem Lösen der neuen Aufgabe lesen mussten (siehe Details unter 5.3). Für die Schülerinnen und Schüler aus den Interventionsgruppen bezog sich diese Infokarte auf die erreichten QSs der vorhergehenden Testaufgabe. Die Schülerinnen und Schüler wussten nicht, dass sich die Infokarten in ihrer Form unterschieden und lösten auch zu diesen beiden Zeitpunkten die jeweilige Aufgabe in Einzelarbeit. Bis auf die Schülerinnen und Schüler aus der KG, erhielten alle Schülerinnen und Schüler zu t_2 und zu t_3 jeweils ein AKF.

IG1. Schülerinnen und Schüler (*n*=38) aus dieser Gruppe erhielten zu t_2 und t_3 jeweils ein FB, welches sich auf die erreichten QSs der vorhergehenden Aufgabe bezog.
IG2. Schülerinnen und Schüler (*n*=41) aus dieser Gruppe erhielten zu t_2 und t_3 jeweils ein FF, welches sich auf das erste nicht erreichte QS der vorhergehenden Aufgabe bezog.
IG3. Schülerinnen und Schüler (*n*=40) aus dieser Gruppe erhielten zu t_2 und t_3 jeweils gleichzeitig ein FB und ein FF.
KG. Schülerinnen und Schüler (*n*=30) aus dieser Gruppe erhielten aus motivationalen Gründen zu t_2 und t_3 auch eine Infokarte. Diese bezog sich jeweils auf den Inhalt der vorhergehenden Aufgabe, hatte jedoch keinen Bezug zum Kompetenzstufenmodell und den QSs.

8.6.3 Intervention mit Infokarten

Das AKF wurden den Schülerinnen und Schüler auf Infokarten vor dem Lösen der zweiten (zu t_2) und dritten (zu t_3) Testaufgabe ausgeteilt. Den Schülerinnen und Schüler wurden jeweils 2 Minuten Zeit gelassen, die Karten zu lesen. Auf allen Karten befanden sich jeweils Comic-artige Ninjas (weiblich/männlich im gleichen Verhältnis stehend), die von Nathan Schreiber (siehe scienceninjas.com) entworfen wurden. Bei den FB-Karten (auch *Defense-Karten* genannt), wurde den Schülerinnen und Schülern mitgeteilt, welche QSs sie in der vorhergehend gelösten Testaufgabe erreicht hatten.

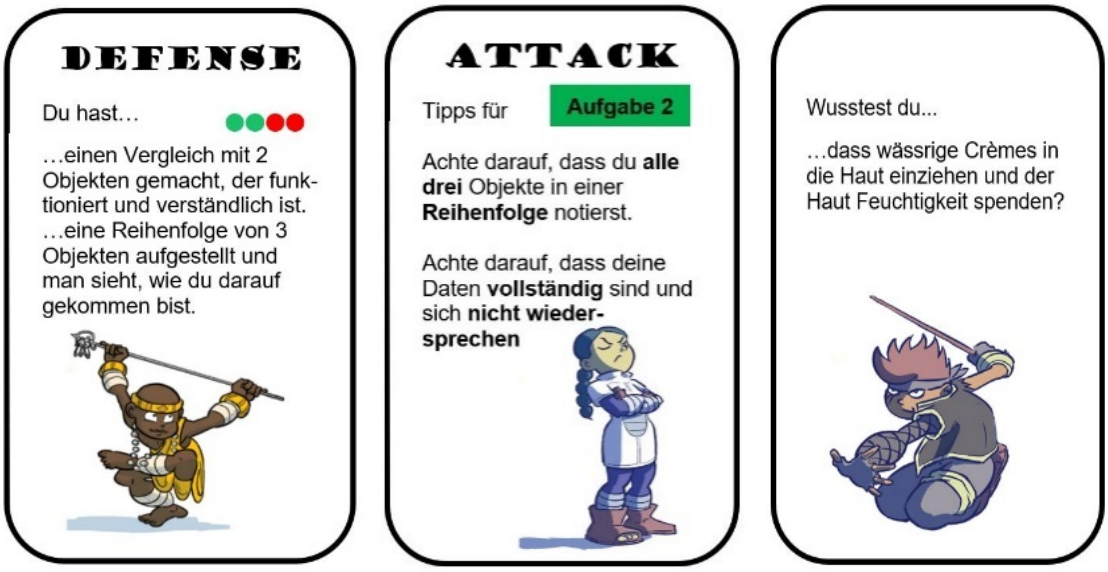


Abb. 8.4 Feeding back (*Defense*), feeding forward (*Attack*) sowie eine Kontrollgruppen-Infokarte aus der Interventionsstudie. Die Figuren auf den Infokarten stammen von Nathan Schreiber (www.scienceninjas.com) und sind urheberrechtlich geschützt.

Abbildung 8.4 zeigt eine Defense-Karte für Schülerinnen und Schüler, welche in der vorhergehenden Testaufgabe QS1 und QS2 erreicht hatten. Des Weiteren wurde durch grüne und rote Punkte auf diesen Karten angegeben, wie viele QSs in der Testaufgabe insgesamt erreicht wurden. Die FF-Karten (*Attack-Karten* genannt) bestanden aus Lernhilfen zum QS, welches, laut Kompetenzstufenmodell, als nächstes erreicht werden sollte (Scaffolding-Maßnahme *hints* bei van de Pol et al., 2010). Abbildung 8.4 zeigt eine Attack-Karte für Schülerinnen und Schüler, welche eine Lernhilfe zum zweiten QS erhielten. Schülerinnen und Schüler aus der IG3 (gleichzeitig FB + FF) erhielten eine Infokarte, welche beidseitig bedruckt war (Defense-Information + Attack-Information). Schülerinnen und Schüler aus der KG erhielten Informationen zum Kontext der vorhergehenden Testaufgabe, welche für die hier untersuchte Kompetenzentwicklung nicht relevant waren (siehe Beispiel in Abb. 8.4). Tabelle 8.2 fasst alle möglichen FBs und FFs aus der Studie zusammen. Mit dieser Vorgehensweise erhielten Schülerinnen und Schüler aus den Gruppen IG1 und IG3, welche viele QSs erreichten, deutlich mehr Text auf den FB-Karten, als Schülerinnen und Schüler, welche nur wenige QSs erreichten. Den Schülerinnen und Schüler aus IG2 und IG3, welche alle QS erreichten hatten, wurden keine FF Karten ausgeteilt, jedoch wurde ihnen jeweils mitgeteilt, dass sie bei der letzten Aufgabe alle geprüften Kriterien erfüllt hatten.

Alle Infokarten wurden in einer Präpilotierung (n = 70, Klasse 8, SekB/C, wobei > 60 % der Schülerinnen und Schüler mit Deutsch als Fremdsprache) ein erstes Mal eingesetzt. Diese Präpilotierung hatte ein ähnliches Design, jedoch gab es nur eine Interventionsgruppe und alle Schülerinnen und Schüler lösten die 4 Testaufgaben in Dyaden. In der Interventionsphase erhielten 18 Schülerpaare vor dem Lösen der zweiten und dritten Testaufgabe Attack- und Defense-Karten gleichzeitig, 17 weitere Paare erhielten keine Infokarten. Erste Ergebnisse aus dieser Studie zeigten, dass die Schülerpaare mit Infokarten eine günstigere Entwicklung ihrer experimentellen Kompetenz zeigten.

In der hier vorliegenden Interventionsstudie wurde allen Schülerinnen und Schülern nach dem Lösen der zweiten und dritten Testaufgabe jeweils ein Fragebogen zur *Wahrnehmung der Unterstützung* bestehend aus 9 Items ausgeteilt, welcher nach dem *Nutzen (3)*, der *Akzeptanz (3)* und der *Fairness (3)* des Unterstützungsangebots fragte (siehe Tabelle 8.3, vierstufiges Antwortformat: 1 – *stimmt überhaupt nicht* bis 4 – *stimmt völlig*).

Tab. 8.2 Mögliches feeding back & feeding forward beim effektbasierten Vergleichen (aus Hild et al., 2018c).

Qualitätsstandard	Feeding back	Feeding forward
kein QS erreicht	Du hast beim letzten Mal keine Punkte erreicht.	Bei Aufgabe 1: Achte darauf, dass deine Daten sich nicht widersprechen. Achte darauf, immer denselben Effekt zu erzeugen.
QS1 erreicht	Du hast einen Vergleich mit 2 Objekten gemacht, der funktioniert und verständlich ist.	Bei Aufgabe 2: Achte darauf, dass du *alle drei* Objekte in einer *Reihenfolge* notierst. Achte darauf, dass deine Daten *vollständig* sind und sich *nicht widersprechen*.
QS2 erreicht	Du hast eine Reihenfolge von 3 Objekten aufgestellt und man sieht, wie du darauf gekommen bist.	Allgemein: Überlege dir, welche *Bedingungen* beim Vergleichen *gleichbleiben* müssen. Achte darauf, dass du immer *Gleiches mit Gleichem* vergleichst.
QS3 erreicht	Du hast faire Vergleiche gemacht.	Bei Aufgabe 3: Achte darauf, dass deine Daten zeigen, welche zwei Objekte *ähnlicher* sind. Achte darauf, dass du *alle* Objekte vergleichst.
QS4 erreicht	Du hast herausgefunden, welche Objekte am ähnlichsten sind.	

Mittels einfaktorieller Varianzanalysen konnte bezogen auf die *Wahrnehmung der Unterstützung* ein Unterschied zwischen der Kontrollgruppe und der Gesamtheit der Interventionsgruppen festgestellt werden. Die beiden Gruppen unterschieden sich signifikant (F (1,147) = 9.368, p = .003, η^2 = .060). Hingegen unterschieden sich die drei Interventionsgruppen bezogen auf die gleiche Variable nicht voneinander *(*F (2,116) = 1.843, p = .163, η^2 = .031). Die gleichen Ergebnisse wurden auch nach t_3 festgestellt. Dies lässt vermuten, dass die *Wahrnehmung der Unterstützung* für die Schülerinnen und Schüler aus den drei Interventionsgruppen ähnlich war.

Tab. 8.3 Mittelwerte & Reliabilität der Variable *Wahrnehmung der Unterstützung* (nach t_2)

Beschreibung (Anzahl items)	**Beispielitem**	***M*/*SD***	**α**	**Quelle**
Wahrnehmung der Unterstützung		3.00 / 0.55	.788	Strijbos Pat-El & Narciss 2010
- Nutzen des Angebots (3)	*Die Rückmeldungen konnte ich brauchen.*	2.83 / 0.75		
- Akzeptanz des Angebots (3)	*Mit den Rückmeldungen war ich nicht einverstanden.*	3.20 / 0.57		
- Fairness des Angebots (3)	*Die Rückmeldungen fand ich fair.*	2.97 / 0.60		

Anmerkungen: M = Mittelwert, *SD* = Standardabweichung, α = Cronbach Alpha, *n = 149*

8.6.4 Instrumente

Experimentelle Kompetenz

Die experimentelle Kompetenz wurde mit Hilfe eines standardisierten Manuals aus dem Projekt ExKoNawi (z. B. Gut et al., 2017) anhand der abgegebenen Protokolle eruiert. Insgesamt wurden 10 Kriterien dichotom kodiert (1 = Kriterium erreicht, 0 = Kriterium nicht erreicht). Schülerinnen und Schüler erreichten pro Beurteilung ein Summenscore *K* von 0 bis 10 Punkten. Für das ausgeteilte AKF wurden aus den 10 Kriterien 4 Qualitätsstandards (1 bis 3 Kriterien pro QS) gebildet (Abb. 8.2). Hierbei galt, dass ein QS als *erreicht* eingestuft wurde, wenn mindestens 2/3 der verwendeten Kriterien im Protokoll erfüllt waren (Gut et al., 2014).

Randomisierung und Überprüfung

Vorgängig zur Interventionsstudie wurde überprüft, ob die Leistungen von Schülerinnen und Schüler aus tiefen Jahrgängen und leistungsschwachen Klassen unabhängig von der zu lösenden Aufgabe und verallgemeinerbar sind (siehe Details bei Hild et al., 2018a, sowie Hild et al., 2018c). Alle 4 hands-on Aufgaben konnten als Paralleltests angesehen werden. Dies wurde auch im Laufe der hier vorgestellten Studie für die Gesamtstichprobe nochmals überprüft (siehe Abb. 8.3). Bei der Gruppenzuteilung wurde darauf geachtet, dass sich Schülerinnen und Schüler mit gleicher Leistung (gleichen erreichten QS) in der Testaufgabe t_1 zufällig auf die vier Gruppen verteilten. Zur Untersuchung der Fragestellungen war eine erfolgreiche Randomisierung unabdingbar. Vor der ersten Aufgabe wurde den Schülerinnen und Schüler

nebst einem Fragebogen mit demographischen Items ein weiterer Fragebogen ausgeteilt, welcher ihre *Lernmotivation im Natur und Technik (NT) Unterricht* mittels der Variablen *Kompetenzerleben im NT Unterricht* und *Interesse im NT Unterricht* abfragte (vierstufiges Antwortformat: 1 – *stimmt überhaupt nicht* bis 4 – *stimmt völlig*). Die Mittelwerte und Reliabilitäten der beiden Variablen finden sich in Tabelle 8.4. Um gleiche Ausgangsbedingungen in den vier Gruppen sicherzustellen, wurde im Nachhinein mittels einfaktorieller Varianzanalysen überprüft, ob sich im Sinne einer gelungenen Randomisierung die vier Gruppen bezüglich dieser Variablen unterschieden (Buff et al., 2010; Rakoczy et al., 2005). Dies mit der zweifachen Begründung: Diese Variablen können generell als Prädiktor von Leistungen angesehen werden (vgl. Eccles 2005 und Wigfield & Eccles 2000). Zweitens kann eine günstige Kompetenzentwicklung, falls keine Unterschiede zwischen den Gruppen bestehen, so eindeutiger mit der Intervention in Verbindung gesetzt werden.

Tab. 8.4 Mittelwerte und Reliabilitäten der Skalen *Komp*etenzerleben und *Interesse.*

Beschreibung (Anzahl items)	**Beispielitem**	***M* / *SD***	**α**	**Quelle**
Kompetenzerleben im NT Unterricht (6)	*In Natur & Technik (NT) bin ich gut.*	2.69 / 0.56	.890 .809	Buff et al., 2010
Interesse im NT Unterricht (6)	*Freiwillig würde ich mich nie mit NT beschäftigen.*	2.87 / 0.56		Rakoczy et al., 2005

Anmerkungen: M = Mittelwert, *SD* = Standardabweichung, α = Cronbach Alpha, *n = 182*

Tab. 8.5 Einfaktorielle Varianzanalysen der gleichen Ausgangsbedingungen.

Variable	**Einfaktorielle Varianzanalyse**
Kompetenzerleben im NT Unterricht – Gruppen (4)	F (3,146) = 1.411, p = .242, η^2 = .029
Interesse im NT Unterricht – Gruppen (4)	F (3,146) = 1.628, p = .186, η^2 = .033

Anmerkungen: F = F-Test (Freiheitsgrade *df*, Stichprobengröβe *N* - *df*), *p* = Signifikanzwert, η^2 = partielles Eta-Quadrat

Tabelle 8.5 zeigt deutlich, dass die vier Gruppen (KG und 3 IGs) keine signifikanten Unterschiede in den, zum Prätest erhobenen Variablen *Kompetenzerleben* und *Interesse* aufwiesen. Auch erwiesen sich die Effektstärken (η^2) für vorhandene Unterschiede als klein. Die Randomisierung scheint entsprechend recht gut gelungen zu sein.

Da der Stichprobendropout hoch war, wurde im Nachhinein überprüft, ob sich die von der Intervention ausgeschlossenen Schülerinnen und Schüler in den beiden Variablen *Kompetenzerleben* und *Interesse* von den übrigen 149 Schülerinnen und Schülern unterschieden, die an der Intervention teilnahmen. Tabelle 8.6 zeigt deutlich, dass keine statistisch oder praktisch relevanten Unterschiede zwischen diesen beiden Gruppen bestanden. Auch bezogen auf das Geschlecht, konnten keine statistisch oder praktisch relevanten Unterschiede festgestellt werden.

Tab. 8.6 Einfaktorielle Varianzanalysen zwischen Dropout und verwendeter Stichprobe

Variable	**Einfaktorielle Varianzanalyse**
Kompetenzerleben im NT Unterricht – Gruppen (2)	F (1,181) = .064, p = .801, $\eta^2 < .001$
Interesse im NT Unterricht – Gruppen (2)	F (1,181) = .380, p = .538, $\eta^2 = .002$

Anmerkungen: F = *F*-Test (Freiheitsgrade *df*, Stichprobengröβe *N* - *df*), *p* = Signifikanzwert, η^2 = partielles Eta-Quadrat

Reliabilitäten

Intrarater-Reliabilität. Für das Aushändigen des AKF (während der Intervention zu den Zeitpunkten t_2 und t_3), mussten die Schülerprotokolle aus den Zeitpunkten t_1 und t_2 schon im Verlauf der Studie ein erstes Mal ausgewertet und die Schülerleistungen beurteilt werden. Die gleiche Person, die dies tat, beurteilte 3 Monate später nochmals alle Testhefte. Damit konnte die Intrarater Reliabilität ermittelt werden. Die Intrarater-Reliabilität wurde auf Ebene der QSs berechnet, da diese für das Erstellen des AKF relevant waren. Die Beurteilung der Intrarater-Reliabilität konnte in allen Fällen als zufriedenstellend eingestuft werden (Tab. 8.7).

Interrater-Reliabilität. Die Schülerprotokolle wurden nach Abschluss der Studie von zwei Personen ausgewertet, wobei eine Person alle Protokolle auswertete (Masterspur). Für die Berechnung der Interrater-Reliabilitäten wurden, nach einer gemeinsamen Trainingsphase 25.5 % (pro Aufgabe je 38 von 185) aller Testhefte doppelt kodiert und die Übereinstimmung aller einzelnen Kriterien aus dem Kodiermanual verglichen. Die Übereinstimmungen lagen bei allen verglichenen Kriterien (1520 Fälle) über 81 % ($\kappa > .87$).

Tab. 8.7 Intrarater-Reliabilität

	QS1	QS2	QS3	QS4
N	370 (=2*185)	370	370	370
p_0	76.6	77.8	91.9	83.5
Cohens κ	.91	.88	< .6 aber Gwet's AC1 > .8	.78
Beurteilung	*zufriedenstellend*	*zufriedenstellend*	*zufriedenstellend*	*zufriedenstellend*

Anmerkungen: N = Anzahl Fälle, p_0 = prozentuale Übereinstimmung, Cohens κ und Gwet's AC1 sind zwei Masse für die Raterüber-einstimmung, bei K < .6, sollte Gwet's AC1 > .8 sein, um eine zufriedenstellende Übereinstimmung zu erreichen (Gwet 2008)

8.6.5 Statistische Analysen zur Überprüfung der Hypothesen

Erste deskriptive Analysen zeigten, dass die abhängige Variable *experimentelle Kompetenz* nicht normalverteilt war. Zur Untersuchung der Fragestellungen wurden deshalb non parametrische Verfahren verwendet. Alle Fragestellungen wurden zweiseitig auf statistische Signifikanz getestet. Zur Prüfung der Fragestellung 1 (Kompetenzentwicklung ohne Lernunterstützung) wurde der Wilcoxon Test verwendet. Zur Prüfung der Fragestellungen 2, 3 und 4 wurde der *U*-Test für Paardifferenzen (Bortz et al., 2000, 279ff) verwendet. Da der *U*-Test für Paardifferenzen keine Standardprozedur in SPSS ist, wurden in SPSS – entsprechend dem Vorgehen in Bortz et al. (2000) – zuerst Differenzwerte zwischen Prä- und Posttest Δ*K* ermittelt. Unterschiede zwischen den zu vergleichenden Gruppen wurden dann mittels eines *U*-Tests geprüft. Aufgrund der Größe der Stichprobe bzw. der einzelnen Gruppen und, weil das klassische Testen von Nullhypothesen auf Signifikanz als „Beleg" bspw. von Unterschieden zwischen Gruppen „seine Tücken hat", sprich, ein statistisch signifikantes Ergebnis noch nichts über dessen praktische Bedeutung aussagt (Bortz et al., 2000; Cohen, 1994; Field, 2009), wurden jeweils die statistische *und* die praktische Signifikanz bzw. die Effektstärke Pearson's *r* ermittelt. Field (2009, 550) stellt dar, wie sich im Falle des *U*-Test die Effektstärke *r* ermitteln lässt ($r = \frac{z}{\sqrt{n}}$). Effektstärken zwischen $0.1 \leq r < 0.3$ werden als klein, zwischen $0.3 \leq r < 0.5$ als mittel und $d \geq .5$ als groß bezeichnet (z. B. Cohen, 1988; Lenhard & Lenhard, 2016). Diese Werte sind allerdings als Richtwerte aufzufassen.

8.7 Ergebnisse

8.7.1 Fragestellung 1

Zeigen Schülerinnen und Schüler auch ohne Lernunterstützung eine Veränderung in der Entwicklung ihrer experimentellen Kompetenz?

Tab. 8.8 Mittelwerte und Standardabweichungen im Prä-/Posttest, sowie im Vergleich (ΔK).

	Prätest (t_1)		**Posttest (t_4)**		**Prä/Post (ΔK)**	
	M	*SD*	*M*	*SD*	*M*	*SD*
gesamt $N = 149$	2.57	2.18	3.46	2.79	0.89	3.08
Kontrollgruppe $n_{KG} = 30$	2.57	2.06	3.03	2.47	0.47	2.52
Interventionsgruppe $n_{IG} = 119$	2.57	2.22	3.57	2.75	1.00	3.19

Anmerkungen: M = Mittelwert, *SD* = Standardabweichung

Tabelle 8.8 zeigt die Mittelwerte und Standardabweichungen der Summenscores *K* für die reduzierte Stichprobe ($N = 149$), die KG ($n_{KG} = 30$) und die IGs ($n_{IG} = 119$, alle Schülerinnen und Schüler mit AKF) im Prä- und Posttest. Im Prätest waren die Werte der beiden Gruppen KG und IGs identisch ($M_{IG} = M_{KG} = 2.57$). Obwohl hier Summenscores verglichen werden, fand die Gruppenzuteilung nicht auf der Ebene dieser Summenscores, sondern auf Ebene der erreichten QSs statt (siehe 8.6.4). Bei den Schülerinnen und Schülern ohne Lernunterstützung (KG) zeigte sich eine günstige Entwicklung ihrer experimentellen Kompetenz, diese erwies sich jedoch als statistisch nicht signifikant und von kleiner praktischen Relevanz (Wilcoxon Test $z = -.963$, $p = .336$, $r = .176$).

8.7.2 Fragestellung 2

Unterscheiden sich die Schülerinnen und Schüler ohne Lernunterstützung in der Entwicklung ihrer experimentellen Kompetenz von der Gesamtgruppe der Schülerinnen und Schüler mit Lernunterstützung?

Tabelle 8.8 zeigt, dass Schülerinnen und Schüler mit Lernunterstützung (IGs) insgesamt in ihrer Kompetenzentwicklung ($\Delta K_{IG} = 1.00$) besser abschneiden als die Schülerinnen und Schüler ohne Lernunterstützung ($\Delta K_{KG} = 0.47$). Die Analyse zeigte allerdings, dass

Schülerinnen und Schüler aus den IGs und der KG sich in der Entwicklung ihrer experimentellen Kompetenz weder statistisch noch praktisch signifikant voneinander unterschieden (U = 1551.0, z = -1.114, p = .264, r = .091). Da die Interventionsgruppe aus drei unterschiedlichen Subgruppen bestand, hätte es sein können, dass sich lediglich eine oder zwei dieser Subgruppen von der Kontrollgruppe unterscheiden (siehe Fragestellung 3).

8.7.3 Fragestellung 3

Unterscheiden sich Schülerinnen und Schüler aus den unterschiedlichen Interventionsgruppen (FB, FF, FB + FF) hinsichtlich ihrer Entwicklung experimenteller Kompetenz je einzeln von den Schülerinnen und Schüler ohne Lernunterstützung?

Tab. 8.9 Mittelwerte und Standardabweichungen im Prä-/Posttest von IG1, IG2, IG3 und KG.

	Prätest (t_1)		Posttest (t_4)		Prä/Post (ΔK)	
	M	*SD*	*M*	*SD*	*M*	*SD*
IG1 (FB) n_{IG1} = 38	2.66	2.37	4.11	3.12	1.45	3.24
IG2 (FF) n_{IG2} = 41	2.56	2.50	3.07	2.47	0.51	3.22
IG3 (FB + FF) n_{IG3} = 40	2.50	2.00	3.58	2.63	1.07	3.16
KG n_{KG} = 30	2.57	2.06	3.03	2.47	0.47	2.52

Anmerkungen: M = Mittelwert, *SD* = Standardabweichung

Tabelle 8.9 zeigt die Mittelwerte und Standardabweichungen der Summenscores für IG1, IG2 und IG3 im Prä- und Posttest. Im Prätest sind die Werte der drei Gruppen nicht ganz identisch (da, wie bereits erwähnt, die Gruppenzuteilung auf Ebene der QSs stattfand). Die durchschnittlichen Summenscores aller einzelnen Gruppen steigen vom Prä- zum Posttest. Zur Untersuchung der Fragestellung 3 wurden die Differenzen der Summenscores ΔK zwischen KG und den jeweiligen IGs auf statistische Signifikanz geprüft und die Effektstärken bestimmt. Von Prä- zum Posttest entwickelte sich die experimentelle Kompetenz der Schülerinnen und Schüler aus der IG1 (ΔK_{IG1} = 1.45) positiver als diejenige der KG (ΔK_{KG} = 0.47). Dieser Unterschied erwies sich als statistisch nicht signifikant, war jedoch von kleiner Effektstärke (U = 435.5, z = -1.675, p = .094, r = .205). Schülerinnen und Schüler aus IG2 (ΔK_{IG2} = 0.51) und der KG unterschieden sich in der Entwicklung ihrer experimentelleren Kompetenz statistisch nicht signifikant voneinander (U = 591.5, z = -.275,

p = .782) und auch die Effektstärke erwies sich mit r = .033 als vernachlässigbar. Bei den Schülerinnen und Schüler aus der IG3 entwickelte sich die experimentelle Kompetenz zwar günstiger als bei denjenigen der Kontrollgruppe (ΔK_{IG3} = 1.07 vs. ΔK_{KG} = 0.47), auch dieser Unterschied erwies sich als statistisch nicht signifikant (U = 524.0, z = -.911, p = .362) und die Effektstärke war eher klein (r = .108).

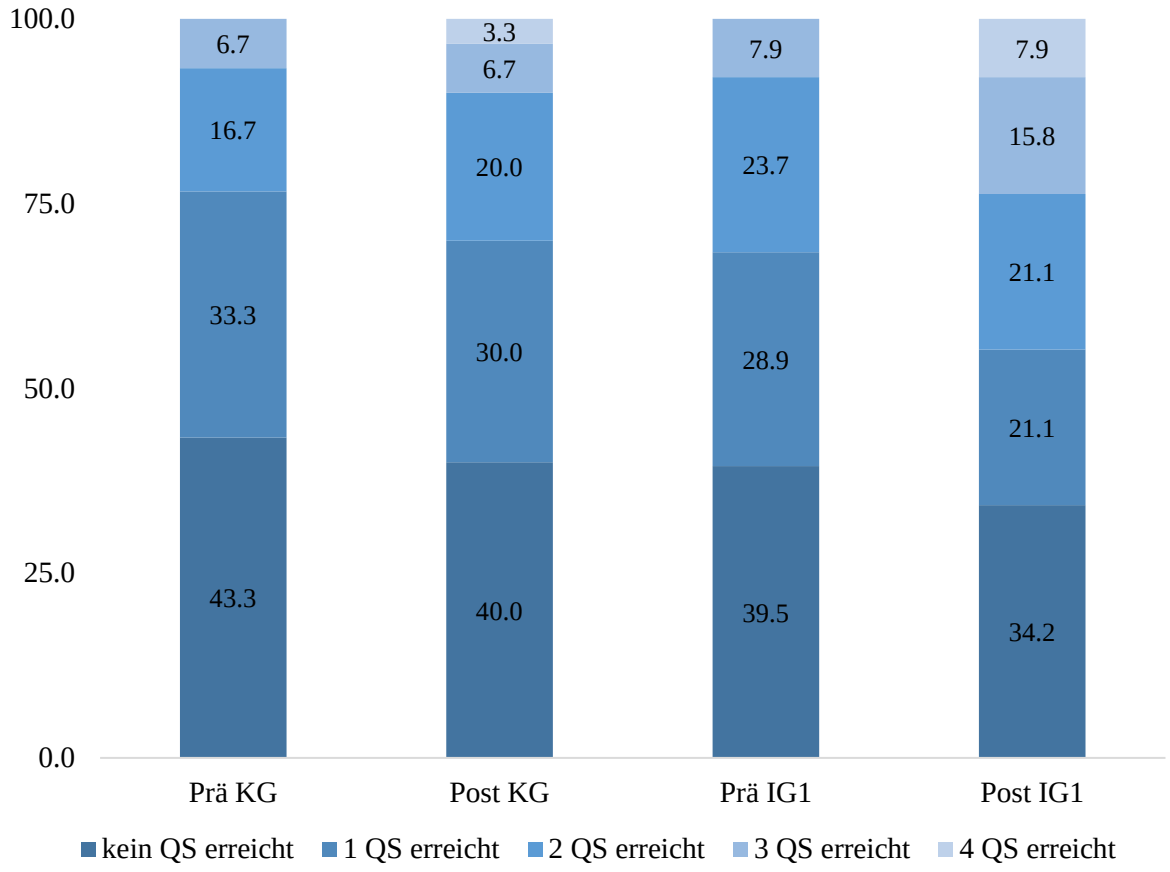


Abb. 8.5 Häufigkeiten erreichter QSs im Prä-/Postvergleich

Da sich die experimentelle Kompetenz von Prä- zum Posttest scheinbar am günstigsten bei Schülerinnen und Schülern der IG1 entwickelte, zeigt Abbildung 8.5 exemplarisch in einem deskriptiven Sinne den Zugewinn an Qualitätsstandards im Vergleich zu Schülerinnen und Schülern aus der Kontrollgruppe. War die Verteilung erreichter QSs zwischen IG1 und KG nach dem Prätest noch relativ ähnlich, so zeigt Abbildung 8.5, dass mehr Schülerinnen und Schüler aus der IG 1 nach dem Posttest drei bis vier Qualitätsstandards erreichen (+15.8 %) als dies in der Kontrollgruppe der Fall (+ 3.3 %) war.

8.4 Fragestellung 4

Unterscheiden sich die Schülerinnen und Schüler mit FB + FF in ihrer Entwicklung experimenteller Kompetenz von Schülerinnen und Schüler mit FB?

Bei den Schülerinnen und Schülern der IG1 (ΔK_{IG1} = 1.45) zeigte sich zwar eine etwas günstigere Entwicklung der experimentellen Kompetenz als bei denjenigen der IG3 (ΔK_{IG3} = 1.07), aber auch dieser Unterschied erwies sich als statistisch nicht signifikant und ist von seiner Effektstärke her praktisch vernachlässigbar (U = 697.50, z = -.628, p = .530, r = .071).

8.8 Diskussion

In der Studie wurde die Wirksamkeit unterschiedlicher adaptiver Feedbackformen auf die Entwicklung experimenteller Kompetenz von Schülerinnen und Schüler aus leistungsschwachen 7. Klassen der Sekundarstufe I untersucht. Wie es scheint, haben sich die validierten hands-on Aufgaben zum effektbasierten Vergleichen basierend auf dem Kompetenzstufenmodell des Projekts ExKoNawi für das Erstellen von AKF geeignet und können als Paralleltests eingesetzt werden. Hinsichtlich der Fragestellungen 1 bis 4 zeigten sich zwar in allen Fällen innerhalb (Fragestellung 1) oder zwischen Gruppen (Fragestellungen 2 bis 4) mehr oder weniger groβe Differenzen in der Entwicklung der experimentellen Kompetenz. Diese deuten darauf hin, dass sich a) die experimentelle Kompetenz auch ohne Lernunterstützung günstig entwickelt (Fragestellung 1), b) deren Entwicklung in den IGs günstiger verläuft als in der KG (Fragestellung 2 und 3) sowie c) sich die IG mit FB positiver entwickelt als die IG mit FF und FB (Fragestellung 4). Allerdings erwiesen sich diese Differenzen *in keinem Falle als statistisch signifikant* ($p < .05$), d.h. sie liessen sich nicht gegen den Zufall absichern.

Ausser Spesen nichts gewesen? Nicht unbedingt, denn das klassische Testen der Nullhypothese auf statistische Signifikanz als „Beleg" für bspw. Unterschiede zwischen Gruppen ist nicht allein von der Gröβe der Unterschiede an sich, sondern u.a. auch von der Stichprobengröβe abhängig (vgl. etwa Bortz et al., 2000; Cohen, 1994). Mit einer (sehr) groβen Stichprobe, bestehen gute Chancen, dass sich auch kleine (oder gar kleinste) Unterschiede als signifikant bzw. überzufällig ($p < .05$) erweisen. Zudem bedeutet ein signifikanter Unterschied noch keineswegs, dass dieser auch von praktischer Bedeutung ist (praktische Signifikanz bzw. Effektstärke). Bei kleinen Stichproben hingegen, kann ein nicht signifikanter Unterschied durchaus von praktischer Bedeutung sein (vgl. etwa Altman & Bland, 1995; Field, 2009). Mit anderen Worten, nicht signifikante Ergebnisse sind nicht per se irrelevant. Effektstärken können im Falle statistisch nicht signifikanter Ergebnisse Hinweise darauf geben, ob hinsichtlich einer Fragestellung oder Hypothese grundsätzlich „Potenzial" vorhanden sein könnte und es sich möglicherweise lohnt, den eingeschlagenen Weg beizubehalten, d.h. die Fragestellung bzw. Hypothese nicht gänzlich zu verwerfen und eine neue Studie – u.a. mit einer gröβeren Stichprobe – ins Auge zu fassen.

Im vorliegenden Falle könnte die Stichprobengröβe in der Tat ein Grund für statistisch nicht signifikante Ergebnisse gewesen sein, worauf eine (a posteriori) Poweranalyse hindeutet (siehe Details bei Faul et al., 2007): Zum Beispiel, um bei der Überprüfung der Fragestellung 3 eine Power von 80% zu erreichen (mit $p \leq 0.1$ und einer Effektstärke von $r = 0.2$) wären mindestens 60 Personen pro Gruppe notwendig gewesen. Dies vor Augen und betrachtet man die geschätzten Effektstärken, so scheinen hinsichtlich AKF vor allem Formen mit feeding back (FB) und in eingeschränkterem Maβe gleichzeitiges feeding back und feeding forward (FB+FF) – im Vergleich zu keiner Lernunterstützung – (möglicherweise) erfolgversprechend, um die experimentelle Kompetenz auch bei Schülerinnen und Schüler aus leistungsschwachen Klassen der Sekundarstufe I (Jahrgangsstufe 7) zu fördern. Davon ausgehend lieβen sich drei Hypothesen für diese Schülerschaft formulieren, die in einer neuen Studie geprüft werden müssten:

Hypothese 1. Feedback vs. kein Feedback. Schülerinnen und Schüler ohne Lernunterstützung unterscheiden sich in der Entwicklung ihrer experimentellen Kompetenz von der Gesamtgruppe von Schülerinnen und Schülern mit Lernunterstützung (FB oder FB + FF). Letztere zeigen eine günstigere Entwicklung.

Hypothese 2. Einfluss der Feedbackform. Schülerinnen und Schüler in den Interventionsgruppen (FB, FB + FF) zeigen *je einzeln* eine günstigere Entwicklung in ihrer experimentellen Kompetenz als Schülerinnen und Schüler ohne Lernunterstützung.

Hypothese 3. Einfluss der Feedbackform. Schülerinnen und Schüler in der Interventionsgruppe (FB) zeigen eine günstigere Entwicklung in ihrer experimentellen Kompetenz als Schülerinnen und Schüler in der Interventionsgruppe (FB+FF).

Mit Blick auf eine neue Studie müssten bzw. könnten zumindest drei Optimierungen vorgenommen werden, um zweifelslos vorhandenen Limitationen der vorliegenden Studie zu begegnen: Erstens müsste die Stichprobe vergröβert werden (vgl. oben). Zweitens wäre zu überlegen, das Treatment zu verstärken, d.h. die Zahl der AKF zu erhöhen (im vorliegenden Falle waren es lediglich zwei). Interessant wäre hier auch die Stärke des Treatments zu variieren, um einen Hinweis darauf zu erhalten, was zwingend nötig erscheint und ab wann kaum mehr groβe Lerneffekte eintreten. Bei einer gröβeren Zahl von AKF wären zudem einzelne Abwesenheiten hinsichtlich der potenziellen Wirksamkeit des Treatments nicht

mehr so problematisch, wie im vorliegenden Falle (vgl. 8.6.1). Eine größere Stichprobe und ein verstärktes Treatment würden es zudem besser erlauben, z.B. *full information maximum likelihood estimation* (FIML) zu verwenden, um dem Problem von Datenausfällen adäquat zu begegnen (Enders, 2010). Drittens wäre zu prüfen, ob es machbar erscheint, die zentrale abhängige Variable (experimentelle Kompetenz) zu den verschiedenen Zeitpunkten latent – via je zwei oder besser drei Indikatoren – zu erheben. Im vorliegenden Falle handelte es sich zu jedem Zeitpunkt um Ein-Item-Messungen mit all der damit verbundenen Reliabilitätsproblematik. Bei latenter Modellierung wäre es möglich, den reliabilitätsbedingten Messfehler auszuschalten (vgl. Geiser, 2012). Genau überlegt werden müsste im Falle einer latenten Erhebung der experimentellen Kompetenz, wie Probleme vermieden werden sollen, die dadurch entstehen könnten, dass einzelne Schülerinnen und Schüler, die zu jedem Zeitpunkt mehrere Aufgaben gelöst haben, möglicherweise für einzelne Aufgaben inhaltlich differente Rückmeldungen erhalten. Gerade bei Schülerinnen und Schüler aus leistungsschwachen Klassen erscheint uns dieser Punkt besonders virulent.
Die vorliegende Studie unterscheidet sich von anderen ähnlichen Studien (vgl. Arnold et al., 2017; Scheuermann, 2018) vor allem darin, dass hier a) Schülerinnen und Schüler leistungsschwacher Klassen, b) methodische Fertigkeiten und Fähigkeiten beim (effektbasierten) Vergleichen im Fokus standen und c) drei häufig diskutierte Formen der Lernunterstützung (FB, FF sowie FB+FF) simultan untersucht wurden.

Die in Abschnitt 8.4.3 und 8.4.4 erwähnten Studien, sowie die hier vorgestellte Intervention, deuten darauf hin, dass die unterschiedlichen Scaffolding-Maßnahmen nicht nur adaptiv, sprich an den Lernstand der Schülerinnen und Schüler, sondern möglicherweise auch stark ans jeweilige Leistungsniveau angepasst sein müssen, um Kompetenzen beim praktisch-naturwissenschaftlichen Arbeiten zu fördern. Bezogen auf die naturwissenschaftsdidaktische Forschung, zeigt diese explorative Studie deutlich, dass, bis dato, solche gruppenspezifischen, differenzierten Untersuchungen fehlen und nötig sind, um tatsächlich Förderangebote bzw. Lernunterstützungen zielgerichtet einsetzen zu können. Hofstetter (2017) redet hier von „ungleichheitssensiblem Unterricht", der, je nach Leistungsniveau, von anderen Diagnose- und Planungstools ausgehen muss (Erkennen von Potenzialen, Orientieren an den Ressourcen), mit dem Ziel, dass alle Schülerinnen und Schülern erfolgreich „weiterkommen im System". Sollte es wirklich so sein, dass, einerseits Schülerinnen und Schüler aus leistungsstarken Klassen durch eine Kombination von FB und FF, andererseits Schülerinnen und Schüler aus leistungsschwachen Klassen durch alleiniges FB die günstigsten

Kompetenzentwicklungen erzielen, hätte dies auch Auswirkungen auf die Ausbildung zukünftiger Lehrpersonen sowie den Unterricht. Die Aussage von Hattie (2008), dass Schülerinnen und Schüler beim allgemeinen Feedback gleichzeitig Informationen zum Lernziel, Lernstand und zu nächsten Lernschritten erhalten sollen, müsste, zumindest für das praktisch-naturwissenschaftliche Arbeiten relativiert werden. Lehrpersonen in leistungsschwachen Klassen könnten aufgefordert werden, bevor sie mit der Planung hinweisgebender Lernunterstützungen (z. B. durch individuelle Förderpläne oder dank einem breiten Angebot an Lernhilfen) beginnen, ihren Schülerinnen und Schülern aufzuzeigen, wo sie Rückstände haben und in welchen Bereichen sie bereits gewisse Standards erreicht haben. Die Spinnennetzmethode (siehe u.a. bei Adamina & Hild, 2019) eignet sich hierfür besonders gut und kann auch in Einzelgesprächen mit den jeweiligen Schülerinnen und Schülern sowie deren Eltern eingesetzt werden.

Leistungsschwache Jugendliche, welche in häufig sehr heterogenen Klassen gemeinsam lernen (siehe 8.6.1), sind für Lehrpersonen eine groβe Herausforderungen (z. B. Krull & Wolfram 2011). Die vorangehenden Ausführungen unterstützten die These von Klieme und Warwas (2011), dass - insbesondere diese Lehrpersonen - viel fachdidaktisches Wissen, Wissen über Diagnostik und Förderung brauchen, damit ein ungleichheitssensibler Unterricht (siehe oben) gewährleistet werden kann. Fachdidaktische Forschung leistet hierzu einen wichtigen Beitrag. Die vorliegende explorative Studie hat Hinweise darauf gegeben, wo Potenzial für künftige fachdidaktische Forschungen im Bereich „Lernunterstützung" bei leistungsschwachen Schülerinnen und Schüler vorhanden sein könnte und hinsichtlich welcher Punkte diesbezügliche neue Untersuchungen optimiert werden könnten.

9 Diskussion und Ausblick

9.1 Zusammenfassung und Diskussion der Ergebnisse

9.1.1. Kompetenzmodellierung und Aufgabenvalidierung

Im Rahmen dieser Dissertation wurden zu 3 Messzeitpunkten resp. mit 3 Stichproben insgesamt 8 Aufgaben (n_{total} = 418 12- bis 15-jährige Schülerinnen und Schüler, siehe auch tab. 6.5 und 6.6) zum experimentellen Problemtyp effektbasiertes Vergleichen verwendet, welche die a priori gesetzten Qualitätsstandards überprüfen sollten (siehe Strukturmodell von ExKoNawi, Abb. 2.1, Abb. 5.1 resp. fig. 6.1). Bei diesem Problemtyp mussten Schülerinnen und Schüler eigenständig Objekte anhand einer gegebenen Eigenschaft bzw. eines erzeugten Effekts experimentell vergleichen (siehe Tab. 2.1, Abschnitte 5.4.1 und 6.5). Bei allen Testaufgaben basierend auf dem Problemtypenansatz aus dem Projekt ExKoNawi lag der Fokus zum größtenteil auf fachmethodischen (z. B. Emulsionen experimentell vergleichen) und z. T. auch überfachlichen (z. B. exaktes und sorgfältiges Arbeiten) Kompetenzen im Bereich des praktisch-naturwissenschaftlichen Arbeitens. Auf Ebene der Teilprozesse wurden bei den Testaufgaben zum effektbasierten Vergleichen vor allem Kompetenzen aus den beiden Bereichen *Durchführung* und *Auswertung* gefördert (siehe Teilprozessansatz bei Emden & Sumfleth, 2012).

Zum Problemtyp effektbasiertes Vergleichen wurden zunächst Grundlagen für die Entwicklung kompetenzorientierter hands-on Testaufgaben beschrieben (Kapitel 5). Der Fokus lag hier vor allem auf dem Erstellen standardisierter Kodiermanuale und dem Arbeiten mit einem Aufgabenstamm, der bei allen Testaufgaben gleich blieb. Zudem wurden Vignetten beschrieben, die die Schülerinnen und Schülern zum Protokollieren animieren sollen (siehe Details in Abschnitt 5.4). Auch wenn diese Testaufgaben als Paralleltests verwendet wurden (in Kapitel 8), muss an dieser Stelle darauf hingewiesen werden, dass einige Kontexte schwieriger waren als andere (siehe vor allem die Unterschiede zwischen den Häufigkeiten der erreichten QS zu Messzeitpunkt II, Abb. 5.3).

Die Studien zur strukturellen Validität zeigten, dass in allen eingesetzten Testaufgaben zu allen Messzeitpunkten eine Lernhierarchie entlang der 4 Qualitätsstandards (QS) aufzufinden ist (Abschnitt 6.7). Bei Schülerinnen und Schülern aus Klassen mit tiefen Anforderungsniveaus wurde der letzte QS jedoch fast nie erreicht (und wurde aus diesem Grund auch nicht abgebildet in den Abbildungen 5.2, 5.3 und 5.4). Bei der Bewertung der Testhefte lagen die Übereinstimmungen zwischen den Prüferinnen und Prüfern alle in einem akzeptablen bis sehr guten Bereich (siehe fig. 6.6), sodass, nebst den Rasch Analysen

(bezogen auf die QS und deren Trennschärfe), ausreichende Argumente für die strukturelle Validität der Aufgaben und letztendlich des Kompetenzmodells geliefert werden konnten.

Zur Beantwortung der zweiten übergeordneten Fragestellung wurde in der zweiten Publikation (Kapitel 6) nebst den Argumenten bzgl. der strukturellen Validität noch nach weiteren Hinweisen einer ausreichenden Validität der Aufgaben gesucht (dies vor allem in den Bereichen *Generalisierbarkeit* und *externe Validität*). In der vorgestellten Publikation wurden 4 Maßnahmen (mit a priori Strukturmodellen arbeiten; standardisierte Aufgaben und Kodiermanuale verwenden; Anzahl Testaufgaben pro SchülerIn erhöhen; mehrere Auswertungs- und Messverfahren gegenüberstellen), die die Validität der Aufgaben erhöhen sollen, im Detail untersucht. Die Studie belegt (Abschnitt 6.7), dass diese Maßnahmen zu einer beachtlichen Erhöhung des Generalisierbarkeitskoeffizienten führen ($\rho^2 = 0.72$). Dem Autor sind im Bereich des praktisch-naturwissenschaftlichen Arbeitens keine anderen naturwissenschaftsdidaktischen Studien bekannt, welche von höheren Koeffizienten berichten (vgl. Abschnitt 6.3.2.1 oder auch Hild et al., 2018b). Dennoch konnte mit diesen Maßnahmen die Personen-Aufgaben Varianz nicht verringert werden. Diese Abhängigkeit bleibt die Achillesferse bei Leistungsmessungen mit experimentellen hands-on Testaufgaben (siehe dazu auch noch 9.2).

Zusammengefasst zeigen die erwähnten Ergebnisse, dass die ersten beiden übergeordneten Fragestellungen in zufriedenstellender Weise beantwortet werden konnten: Die Qualitätsstandards eignen sich zur Beschreibung experimenteller Kompetenzen von Schülerinnen und Schülern der Sekundarstufe I, insbesondere können sie als Stufen einer Lernhierarchie identifiziert werden und sie lassen sich zudem mit dem gewählten Messverfahren (Auswerten von Protokollen) reliabel und valide erfassen. Die Ergebnisse aus den beiden ersten Publikationen dienten als Grundlage für die Konzipierung einer Intervention (Kapitel 8), welche erlaubte, Kompetenzen im Bereich des praktisch-naturwissenschaftlichen Arbeitens zu beurteilen und individuell zu fördern (siehe 9.1.3).

9.1.2 Umsetzung in der Praxis

In der dritten Publikation wurden experimentelle Lerngelegenheiten vorgestellt, welche sich explizit mit dem experimentellen Problemtyp kategoriengeleitetes Beobachten befassen. Die Publikation richtet sich an Lehrpersonen und liefert konkrete Hinweise zum Erstellen kompetenzorientierter Beobachtungsaufträge. Unter anderem wird hier nochmals dargestellt, dass unspezifisch formulierte Beobachtungsaufträge dazu führen können, dass häufig das wahrgenommen wird, was für naturwissenschaftliche Beobachtungen irrelevant ist (siehe Abb. 7.1).

Bei den beiden vorgestellten Lerngelegenheiten (Abb. 7.4 und 7.5) wurden zwei bis drei Phänomene einzeln beobachtet, beschrieben und miteinander verglichen (eine pH-abhängige vs. eine pH-unabhängige Tintensorte; zwei Wasser-in-Öl- vs. eine Öl-in-Wasser-Emulsionen). Die Aufgabenstellung richtet sich dabei entlang der gefundenen Lernhierarchie zu diesem Problemtyp (siehe Abschnitte 2.4.2. und 7.2). Mit Hilfe einer Methodenkarte (Abb. 7.2) können Schülerinnen und Schüler im Bereich ihrer fachmethodischen und überfachlichen Kompetenzen unterstützt werden.

Die dritte übergeordnete Fragestellung konnte hier insofern beantwortet werden, dass die gefundene Lernhierarchie das Erstellen kompetenzorientierter Beobachtungsaufgaben unterstützte, jedoch muss an dieser Stelle ganz klar auch Kritik ausformuliert werden: Einschränkend ist festzuhalten, dass diese Lerngelegenheiten weder auf ihre Lernwirksamkeit (Kompetenzerwerb) überprüft, noch mit anderen Aufgaben bzw. Aufgabenformulierungen verglichen wurden.

9.1.3 Formative Beurteilung und Art der Lernunterstützung

Wie bereits in Abschnitt 2.9 erwähnt, gibt es nur wenige Interventionen, die den Einfluss von adaptiven Fördermaßnahmen auf die experimentellen Kompetenzen von Schülerinnen und Schülern untersucht haben (z. B. Arnold et al., 2017; Scheuermann, 2018; Wollenschläger et al., 2012). Zudem standen in keiner der Studien Schülerinnen und Schüler aus leistungsschwachen Klassen (Kategorie *Grundansprüche)* im Fokus.

Die Interventionsstudie konnte belegen, dass sich das Kompetenzstufenmodell zum experimentellen Problemtyp effektbasiertes Vergleichen für das Erstellen von adaptivem, kompetenzbezogenen Feedback eignet. Die dazugehörigen Testaufgaben konnten in dieser Intervention als Paralleltests eingesetzt werden. Die übergeordnete Fragestellung 4 konnte jedoch nicht vollumfänglich beantwortet werden: So konnte nicht bestätigt werden, dass die eingesetzten adaptiven Fördermaßnahmen (hinweisgebende bzw. rückmeldende

Lernunterstützungen bzw. beides gleichzeitig) zu einer günstigeren Kompetenzentwicklung im Vergleich zu Schülerinnen und Schüler aus einer Kontrollgruppe ohne Fördermaßnahmen führen (siehe 8.7.2). Eventuell sind die Kompetenzbeschreibungen der Qualitätsstandards im problemtypenbasierten Kompetenzmodell zu unspezifisch um Schülerinnen und Schülern beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten zu unterstützen.

Die Ergebnisse zur letzten übergeordneten Fragestellung (5) deuten darauf hin, dass bei Schülerinnen und Schülern aus leistungsschwachen Klassen adaptive Fördermaßnahmen dann wirksam sind, wenn sie bzgl. ihrer erreichten Kompetenzen ein feeding back erhalten (siehe 8.7.3). Hier müssen jedoch noch weitere Interventionsstudien durchgeführt werden, vor allem sollten größere Stichproben verwendet werden.

9.2 Theoretische und praktische Implikationen für die Entwicklung naturwissenschaftlicher Fördermaßnahmen

„In der universitären Lehrerbildung muss es … Ziel sein, Studierende zu befähigen, auf Grundlage geeigneter Diagnoseinstrumente fachliche Lernumgebungen zu gestalten, in denen Lernende in Formen *Offener Differenzierung* … gemeinsam erfolgreich lernen können" (Selter et al., 2017, 12). Diesem Desiderat, welches auch in ähnlicher Form bei Krammer (2009) ausformuliert wurde (siehe Abschnitt 8.2), konnte im Rahmen dieser Dissertation nachgegangen werden. Es wurde aufgezeigt, dass sich Kompetenzstufenmodelle als Planungs- und Diagnosetool eignen, hier im Bereich fachmethodischer und überfachlicher Kompetenzen. Nebst der vorgestellten Interventionsstudie (Kapitel 8), kommen auch weitere Autorinnen und Autoren zu diesem Fazit (z. B. Bernholt, Parchmann & Commons, 2009; Wollenschläger et al., 2012). Ob nun als Fördermaßnahme eher hinweisgebende oder doch lieber rückmeldende Lernangebote eingesetzt werden, hängt wohl mit dem Anforderungsniveau der Stichprobe zusammen: Für Schülerinnen und Schüler aus leistungsschwachen Klassen scheint beim selbstständigen praktisch-naturwissensachaftlichen Arbeiten zusätzliches feeding back zu einer günstigeren Kompetenzentwicklung zu führen - für leistungsstarke Schülerinnen und Schüler ist hinweisgebendes feeding forward gekoppelt an rückmeldenes feeding back wirksamer (z. B. Scheuermann, 2017).

Für die Praxis bedeutet dies, dass Lehrpersonen von Klassen mit leistungsschwachen Schülerinnen und Schülern, bevor sie mit der Planung individueller hinweisgebender Unterstützungen für selbstständiges Lernen (z. B. durch individuelle Förderpläne oder einem breiten Angebot an Scaffolds) beginnen, ihren Schülerinnen und Schülern zuerst oder auch vor allem aufzeigen müssen, in welchen Bereichen sie bereits erste Erfolge erzielt haben und wo ein großes Lernpotenzial besteht.

Bezogen auf die naturwissenschaftsdidaktische Forschung zeigt die letzte Publikation (Kapitel 8), dass gruppenspezifische, differenzierte Untersuchungen und Aussagen nötig sind, um tatsächlich Förderangebote zielgerichtet einsetzen zu können. Zurzeit fehlen weitere naturwissenschaftsdidaktischen (Interventions-)Studien, die den Einsatz von Kompetenzmodellen als Diagnose- und Fördertool bei Schülerinnen und Schülern aus der Kategorie Grundansprüche untersuchen (bzw. aus der Haupt- und Realschule).

9.3 Limitationen der Arbeit

In dieser Arbeit wurden einige Abgrenzungen getroffen (Kompetenzmodellierung, Aufgabenkontexte, Stichprobenwahl, Aspekte der Validierung, Einfluss bestimmter Lernervariablen,…) welche dazu führen, dass die hier getroffen Aussagen nicht verallgemeinerbar sind. Zu den Limitationen der Arbeit sollen folgende Punkte noch einmal kurz diskutiert werden:

- Praktisch-naturwissenschaftliches Arbeiten wird hier nur exemplarisch am Problemtyp effektbasiertes Vergleichen untersucht. Zur Beschreibung des praktisch-naturwissenschaftlichen Arbeitens müssten jedoch weitere (bestenfalls alle möglichen Problemtypen/Teilprozesse) untersucht werden.
- Die eingesetzten Testaufgaben fokussieren Kontexte, welche keinen Lehrplanbezug haben. Auch wenn die Problemstellungen eine gewisse Alltagsnähe vorweisen, wäre es wichtig herauszufinden, schon alleine zu Validierungszwecken, welchen Einfluss das Vorwissen bei Testaufgaben zum effektbasierten Vergleichen hat und inwiefern fachinhaltliche und fachmethodische Kompetenzen wirklich trennscharf untersucht und gefördert werden können. Es fehlt an dieser Stelle sicherlich eine Studie, welche den Zusammenhang zwischen inhaltlichem Vorwissen (auf die untersuchten Kontexte bezogen) und experimenteller Kompetenz (als Variable) untersucht. Bezogen auf den Zusammenhang zwischen strategischem Vorwissen und experimenteller Kompetenz wurden im Projekt ExKoNawi erste Ergebnisse bereits veröffentlicht (siehe dazu Bonetti et al., 2019).
- Die Testaufgaben und letztlich auch das Kompetenzstufenmodell wurden auf verschiedene Aspekte einer gelungenen Validierung überprüft. Im Sinne von Messick (1996) und einer Konstruktvalidität fehlen an dieser Stelle noch weitere Studien, welche sich vor allem mit der kognitiven Validität der Aufgaben sowie weiteren Argumenten einer externen Validität auseinandersetzen (siehe dazu weitere Forschungsdesiderata unter Abschnitt 9.4).
- Die Personen-Aufgaben Varianz ist nach wie vor zu hoch. Dies führt längerfristig dazu, dass das hier untersuchte Testinstrument wohl kaum in einem large-scale Verfahren seine Verwendung finden wird (im Sinne eines Bildungsmonitorings). Einerseits darf bei dem vorgestellten Messinstrument die Anzahl Aufgaben pro Person nicht reduziert werden (sonst würde sich der Generalisierbarkeitskoeffizient wieder verschlechtern), andererseits müssten noch viele weitere Testaufgaben eingesetzt werden um das Konstrukt der experimentellen Kompetenz(en) der Schülerinnen und Schüler zu überprüfen. Dies würde zu einer viel zu langen Testzeit führen und das ganze Instrumentarium wäre zu kostspielig, um groβe Stichproben damit zu testen. Es scheint jedoch auch keine Alternative resp.

andere Testverfahren zu geben, welche das praktisch-naturwissenschaftliches Arbeiten in seiner Vielfalt abdecken. Hier wird man sich auch längerfristig auf bestimmte experimentelle Kompetenzen fokussieren müssen (wie dies schon jetzt getan wird: Die meisten Modelle fokussieren dabei das Experimentieren als hypothesengeleitetes Forschen durch Manipulation der unabhängigen Variable).

- Die hier vorgestellten Lerngelegenheiten zum kategoriengeleiteten Beobachten (Kapitel 7) müssten unbedingt noch auf ihre Lernwirksamkeit überprüft werden. In der aktuellen Form fehlen wichtige Hinweise darüber, ob mit diesen Aufgaben bestimmte Kompetenzen (hier vor allem die a priori gesetzten QS) im Bereich des Beobachtens wirklich stärker unterstützt bzw. gefördert werden können. Auch hier muss man davon ausgehen, dass die Fördermaßnahmen von der Stichprobe und (hier sicherlich auch) den gewählten Kontexten abhängen.
- Ein Aspekt, welcher in der Interventionsstudie nicht breit genug diskutiert wurde, ist die Lernmotivation der Schülerinnen und Schüler. Nebst der Frage, inwiefern die formative Beurteilung einen Einfluss auf die Kompetenzentwicklung hat, stellt sich die Frage, inwiefern die formative Beurteilung einen Einfluss auf die Lernmotivation der Schülerinnen und Schüler hat. Die empirische Befundlage zu dieser Fragestellung ist eher spärlich (z. B. Rönnebeck, Bernholt & Ropohl, 2016). Nach Deci und Ryan (1991) kann davon ausgegangen werden, dass, auf das Kompetenzerleben gerichtete Feedbacks einen positiven Einfluss auf die Entwicklung einer *aktuellen Lernmotivation* haben. Dies ist jedoch nicht immer der Fall (Yin et al., 2008). Bei Harks et al. (2014b) konnte eine positive indirekte Wirkung kompetenzbezogener Feedbacks auf die beiden motivationsbezogenen Variablen Selbstwirksamkeit und (intrinsisches) Interesse nachgewiesen werden. Vollmeyer und Rheinberg (2005) konnten zeigen, dass Schülerinnen und Schüler, die ein Feedback erwarten, von Beginn an, bessere Strategien nutzen. Des Weiteren wurde gezeigt, dass positiv formuliertes Feedback (im Klassenzimmer) eine Steigerung der, hingegen negativ formuliertes Feedback keinen Einfluss auf die intrinsische Motivation hat (Raaijmakers et al., 2017; Rakoczy et al., 2008). Die *wahrgenommene Unterstützung des Feedbacks* als vermittelnde Variable der Feedbackeffekte auf den Lernzuwachs sowie auf die intrinsische Motivation wurde u. a. bei Harks et al. (2014a) und bei Scheuermann (2017) aufgezeigt und sollte in einer nächsten Studie auch mitgemessen werden.
- Ein letzter Aspekt, welcher unbedingt noch hervorgehoben werden muss, ist die Unterpowerung der Interventionsstudie (post hoc Power Analyse). Die in Kapitel 8 vorgestellte Publikation wurde, nach einem dropout von 36 Schülerinnen und Schülern,

mit einer reduzierten Stichprobe von 149 Schülerinnen und Schülern durchgeführt. Mit 3 Interventionsgruppen und einer Kontrollgruppe führte dies zu sehr kleinen Vergleichsgruppen. Da die Stichprobengröβe einen Einfluss auf die Prüfung statistischer Signifikanz hat und die Effekte kleiner waren als erwartet, konnten die Fragestellungen der Interventionsstudie zum grössten Teil nicht gründlich genug beantwortet werden (zum Teil war nur praktische Signifikanz gegeben). Der Pilotcharakter der Interventionsstudie muss an dieser Stelle unbedingt nochmals erwähnt werden.

9.4 Ausblick auf relevante Forschungsdesiderata und Entwicklungsmöglichkeiten

Zurzeit laufen zwei Dissertationsprojekte in denen die hier vorgestellte problemtypenbasierte Kompetenzmodellierung untersucht wird. Angela Bonetti untersucht im Rahmen der strukturellen Validität des Modells, inwieweit drei unterschiedliche Problemtypen (skalenbasiertes Messen, effektbasiertes Vergleichen und fragengeleitetes Untersuchen), die unterschiedliche Anforderungen an transferfähiges Strategiewissen stellen, eigenständige Konstrukte darstellen und inwiefern die Experimentierleistung mit Strategiewissen, sprachlichen und kognitiven Fähigkeiten korrelieren (Bonetti, Metzger & Gut, 2017). Livia Murer untersucht exemplarisch am Problemtyp skalenbasiertes Messen die Passung der intendierten kognitiven Prozesse bei der Kompetenzerfassung zum postulierten theoretischen Kompetenzmodell (kognitive Validität) und vergleicht unterschiedliche Messverfahren miteinander (Schülerprotokolle, Videos und Interviews).

Die 24 hands-on Testaufgaben, die im Laufe des Projekts ExKoNawi entwickelt und validiert wurden, sollten auch weiterhin, ähnlich wie in der hier vorgestellten Studie (Kapitel 8) als Paralleltests bei Interventionsstudien eingesetzt werden. Zurzeit existieren wenige validierte Testinstrumente, welche vor allem fachmethodische und überfachliche Kompetenzen im Bereich des praktisch-naturwissenschaftlichen Arbeitens im Unterricht der Sekundarstufe I abdecken. Es wäre sinnvoll, wenn weitere Interventionsstudien mit den hier erwähnten Testaufgaben folgen könnten, nicht zuletzt auch im Sinne einer konsequentiellen Validität (siehe Tab. 2.4).

Obwohl in der vorgestellten Interventionsstudie das (rückmeldende) feeding back als erfolgreiche Fördermaßnahme für Schülerinnen und Schüler aus leistungsschwachen Klassen (hier Jahrgangsstufe 7) identifiziert werden konnte, fehlen hier weitere Studien, welche die (vor allem sprachliche) Qualität des Angebots, sowie den Einfluss auf die (aktuelle) Lernmotivation untersuchen.

Bei Schülerinnen und Schüler aus leistungsstarken Klassen könnten eventuell nebst hinweisgebendem feeding forward auch Selbstbeurteilungsbögen (Schreiber & Theyßen, 2019) oder Lösungsbeispiele (Koenen, Kölbach, Emden & Sumfleth, 2014) zu ähnlichen Erfolgen führen. Diese unterschiedlichen Fördermaßnahmen müssten jedoch nochmals miteinander verglichen werden.

Allgemein zeigt sich immer deutlicher in der Naturwissenschaftsdidaktik, dass eine Vielfalt an Aufgaben und eine Vielfalt an Fördertools noch lange nicht zu einer gewissen Adaptivität beim Lernen (hier beim praktisch-naturwissenschaftlichen Arbeiten) führen müssen. Zudem wurden gruppenspezifische Unterschiede (Kategorie Grundansprüche vs. Kategorie

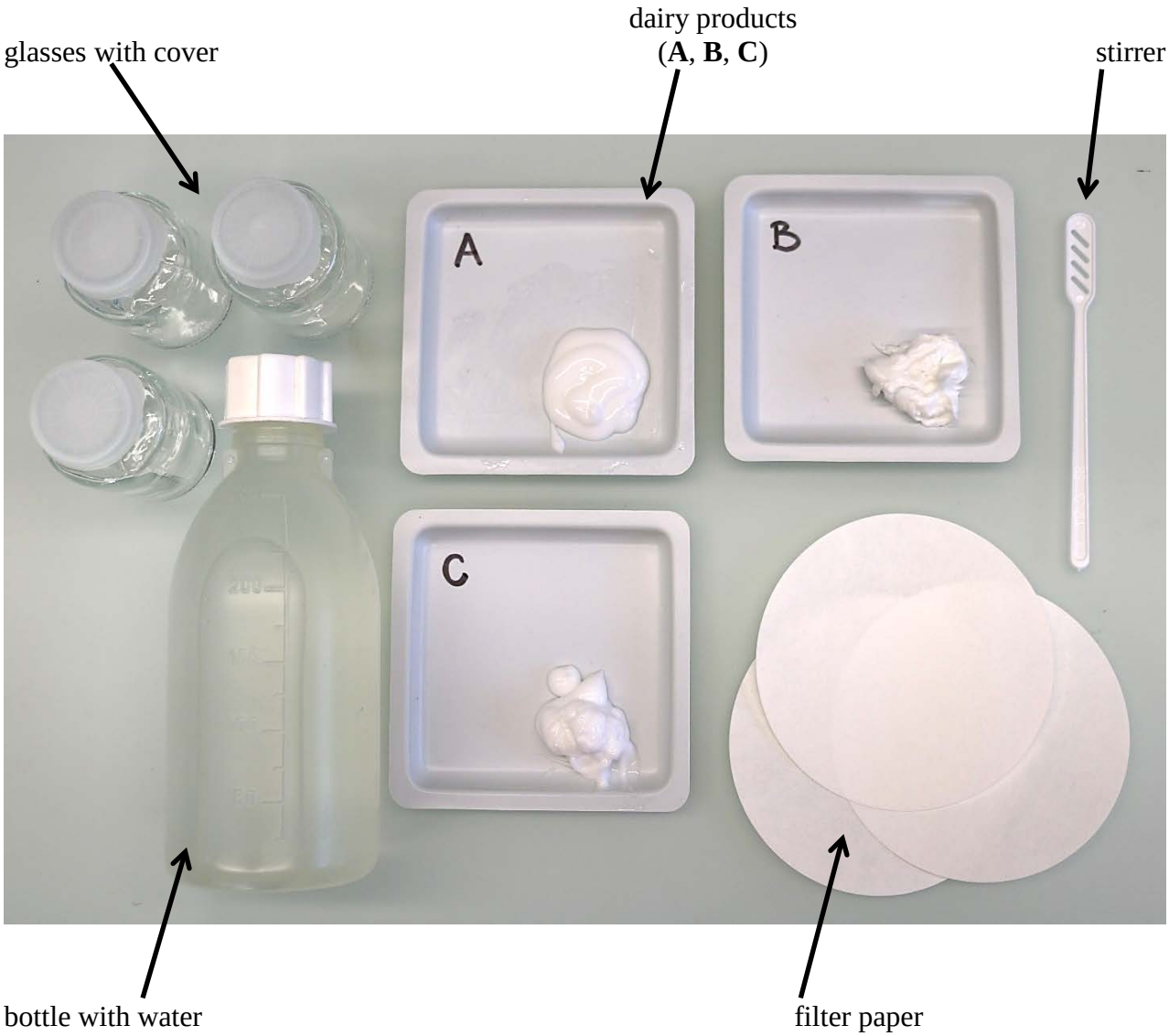
erweiterte Ansprüche) bis dato nicht oder nur am Rande beachtet. Gerade in diesem Bereich (und auch im Vergleich mit den Fachdidaktiken der Hauptfächer Deutsch und Mathematik) hat die Didaktik der Naturwissenschaften einen erhöhten Nachholbedarf.

Abschließen möchte ich diese Arbeit mit einer, so finde ich passenden Aussage von Tweney: „Faraday never discovered anything with a single experiment, and neither, I suspect, did anyone else (Tweney, 1990, 475)".

Anhang

Dairy products

Material



Problem

Olivia and Otto must find out, which dairy product has the best water properties.

They don’t know how to do it.
Find it out for Olivia & Otto by solving the following **4 tasks**.

Background

Dairy products are made of water and fat. They make different spots on paper or behave differently if you mix them with water.

Task ①

Compare *A* with *B*.
Find out which dairy product has the best water properties.

➢ What did you find out? Mark the right answer:

☐ **A** has better water properties than **B**.

☐ **B** has a better water properties than **A**.

Olivia & Otto have found a different result. They want to know, how you managed to get yours. Describe and sketch which experiments and observations you did.

➢ Explain it to Olivia & Otto in a way that they can do the comparative investigation by themselves.

Task ②

Continue by including *C*.

➢ Find out a sequence for dairy products **A**, **B** and **C**.
Start with the dairy product, that has the best water properties.

…… …… ……

Olivia & Otto have found a different result. They want to know, how you managed to get yours. Describe and sketch which experiments and observations you did.

- Explain it to Olivia & Otto in a way that they can do the comparative investigation by themselves.

Task ③

How did you ensure that your comparative investigations were fair?

Task ④

Find out, which two dairy products are most similar, if you examine the water properties. If needed, make further comparisons to solve this task.

- What did you find out? Mark the right answer:

  ☐ **A** and **B** are most similar.

  ☐ **A** and **C** are most similar

  ☐ **B** and **C** are most similar.

Olivia & Otto have found a different result. They want to know, how you managed to get yours. Describe and sketch which experiments and observations you did.

- Explain it to Olivia & Otto in a way that they can do the comparative investigation by themselves.

- Mark the material you have used in your comparative investigations.

  ☐ filter paper

  ☐ glasses with cover

  ☐ Stirrer

Literaturverzeichnis

Adamina, M. & Hild, P. (2019). Mit Lernaufgaben Kompetenzen fördern. In P. Labudde & S. Metzger (Hrsg.), *Fachdidaktik Naturwissenschaften, 1.–9. Schuljahr*, 3te Auflage. Bern: Haupt Verlag.

Allal, L. (2010). Assessment and the regulation of learning. *International encyclopedia of education, 3*, 348–352.

Altman, D. G. & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ 311*, 485.

Arnold, J., Wellnitz, N. & Mayer, J. (2010). Beschreibung und Messung von Beobachtungskompetenz bei Schülerinnen und Schülern der Sekundarstufe I. In D. Krüger, A. Upmeier zu Belzen & S. Nitz (Hrsg.), *Erkenntnisweg Biologiedidaktik* (S. 7–22). Beiträge auf der 12. Frühjahrsschule der Fachsektion Didaktik der Biologie im Verband Biologie, Biowissenschaften und Biomedizin in Deutschland (VBIO) in Neumünster.

Arnold, J. (2015). *Die Wirksamkeit von Lernunterstützungen beim Forschenden Lernen: Eine Interventionsstudie zur Förderung des Wissenschaftlichen Denkens in der gymnasialen Oberstufe.* Dissertation IPN Kiel.

Arnold, J., Kremer, K. & Mayer, J. (2017). Scaffolding beim Forschenden Lernen: Eine empirische Untersuchung zur Wirkung von Lernunterstützungen. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*(1), 21-37.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181-214.

von Aufschnaiter, C., Selter, C. & Michaelis, J. (2017). Nutzung von Vignetten zur Entwicklung von Diagnose- und Förderkompetenzen – Konzeptionelle Überlegungen und Beispiele aus der MINT-Lehrerbildung. In C. Selter, S. Hußmann, C. Hößle, C. Knipping, K. Lengnink & J. Michaelis (Hrsg.), *Diagnose und Förderung heterogener Lerngruppen. Theorien, Konzepte und Beispiele aus der MINT-Lehrerbildung* (S.85–106). Münster: Waxmann.

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C. & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science*, 33, 381–412.

Bangert-Drowns, R. L., Kulik, C.-L. C, Kulik, J. A. & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213.

Barzel, B., Reinhoffer, B. & Schrenk, M. (2012). Das Experimentieren im Unterricht. In W. Rieß, M. Wirtz, B. Barzel & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht* (S. 103-127). Münster: Waxmann.

Baumert, J., Bos, W. & Lehmann, R. H. (2000). *TIMSS/II & TIMSS/III – Dritte internationale Mathematik- und Naturwissenschaftsstudie*. Opladen: Leske & Budrich.

Bell, B. & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science education*, *85*(5), 536–553.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Bernholt, S., Parchmann, I. & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. *Zeitschrift für Didaktik der Naturwissenschaften*, *15*, 219–245.

Berry, D. C. (1991). The role of action in implicit learning. *The Quarterley Journal of Experimental Psychology*, *43A*(4).

Bildungsdirektion Kanton Zürich (2018). *Die Schulen im Kanton Zürich 2017/2018.* https://www.bista.zh-.ch/_pub/downloads/Schulen_Kt_ZH_2017_18.pdf (letzter Zugriff am 14.6.2019)

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment in education: principles, policy & practice*, *5*(1), 7–74.

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*, 3–13.

Blömeke, S., Risse, J., Müller, C., Eichler, D. & Schulz, W. (2006). Analyse der Qualität von Aufgaben aus didaktischer Sicht. Ein allgemeines Modell und seine exemplarische Umsetzung im Unterrichtsfach Mathematik. *Unterrichtswissenschaft*, *4*, 330–356.

Bloom, B. (1984). The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6).

Bonetti, A., Metzger, S. & Gut, C. (2017). Validierung des ExKoNawi-Modells (Experimentelle Kompetenzen in den Naturwissenschaften). In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 336–339). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Zürich 2016.

Bonetti, A., Gut, C., Metzger, S. & Walpuski, M. (2019). Performanz beim Experimentieren mit und ohne Experimentiermaterial. In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe* (S. 73–76). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Kiel 2018.

Börlin, J. (2012). *Das Experiment als Lerngelegenheit. Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität.* Dissertation. Berlin: Logos.

Bortz, J., Lienert, G. A. & Böhnke, K. (2000). *Verteilungsfreie Methoden der Biostatistik, 2. Aufl*. Berlin: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler, 4. Aufl*. Berlin: Springer.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Philips (ed.), *Technical issues in large-scale performance assessment*. Washington D.C.: National Center for Education Statistics.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353.

Bruner, J. S. (1962). *The process of education.* Harvard University Press.

Buff, A., Dinkelmann, I., Steiner, E. & Reusser, K. (2010). *Transition. Elterliche Unterstützung und motivational-affektive Entwicklung beim Übertritt in die Sekundarstufe I: Detaillierte Dokumentation der quantitativen Erhebungen auf situationsspezifischer Ebene Dezember 2008 – März 2009*. Zürich: Pädagogische Hochschule Zürich & Institut für Erziehungswissenschaft, Universität Zürich.

Chen, O., Kalyuga, S. & Sweller, J. (2016). When instructional guidance is needed. *The Educational and Developmental Psychologist, 33*(2), 149–162.

Chiu, M. H., Chou, C. C. & Liu, C. J. (2002). Dynamic processes of conceptual change: analysis of constructing mental models of chemical equilibrium. *Journal of Research in Science Teaching,* 39, 688–712.

Coelho, S. M. & Séré, M.-G. (1998). Pupils' reasoning and practice during hands-on activities in the measurement phase. *RSTE,* 16(1), 79–96.

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist, 49,* 997–1003.

Cohen, J. (2008). Explaining psychological statistics (3rd ed.). New York: John Wiley & Sons.

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson, T. L., Draney, K. & Adams, K. M. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement, 9*(2).

Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16*(2).

Cronbach, L. J., Linn, R. L., Brennan, R. L. & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*(3).

DeBoer, G. E. (2000). Scientific literacy: another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37*(6), 582–601.

Deci, E. L. & Ryan, R. M. (1991). A motivational approach to self integration in personality. In R. Dienstbier (Hrsg.), *Nebraska symposium on motivation, Vol 38, Perspectives on motivation* (pp. 237–288). Lincoln, NE: University of Nebraska Press.

D-EDK – Deutschschweizer Erziehungsdirektoren-Konferenz (2015). *Lehrplan 21. Grundlagen.* https://nw.lehrplan.ch/index.php?code=e|200|1 (letzter Zugriff am 23.5.2019)

Dewey, J. (1910). *How we think.* Boston, New York, Chicago: D. C. Heath & Co. Publishers.

Duit, R., Gropengießer, H. & Stäudel, L. (2007). *Naturwissenschaftliches Arbeiten, Unterricht und Material 5–10, 2. Auflage* (S. 10–11). Seelze-Velber: Friedrich Verlag.

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (eds.), *Handbook of competence and motivation* (pp. 105–121). New York: Guilford.

EDK. Schweizerische Konferenz der kantonalen Erziehungsdirektoren (2011). Grundkompetenzen für die Naturwissenschaften. Nationale Bildungsstandards. https://edudoc.ch/record/96787/files/-grundkomp_nawi_d.pdf (letzter Zugriff am 23.05.2019)

Ellett, C. (1986). Conceptualizing the Study of Learning Environments. In B. Fraser (ed.), *The Study of Learning Environments, Vol I* (p. 34). Salem: Assessment Research.

Emden, M., Bewersdorff, A., & Baur, A. (2019). Kann Experimentieren in der Schule bilden? *Zeitschrift für Pädagogik, 5*, 710.

Enders, C. T. (2010). *Applied missing data analysis.* New York: The Guilford.

Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I.* Dissertation. Berlin: Logos.

Emden, M. & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung des experimentellen Arbeitens. Zur Eignung einer Protokollmethode zur Bewertung von Experimentierprozessen. *Der mathematische und naturwissenschaftliche Unterricht, 65*(2), 68–74.

Emden, M., Bewersdorff, A., & Baur, A. (2019). Kann Experimentieren in der Schule bilden? *Zeitschrift für Pädagogik, 5*, 710.

Enders, C. T. (2010). *Applied missing data analysis.* New York: The Guilford.

Erickson, G. (1994). Pupils' understanding of magnetism in a practical assessment context: the relationship between content, process and progression. In P. Fensham, G. Richard & R. White (eds.), *The content of science.* London: Falmer.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Fay, M. E., Grove, N. P., Towns, M. H. & Bretz, S. L. (2007). A rubric to characterize inquiry in the undergraduate chemistry laboratory. *Chemistry Education Research and Practice*, 8(29), 212-219.

Fechner, S. (2009). *Effects of context-oriented learning on student interest and achievement in chemistry education.* Dissertation. Berlin: Logos.

Felouzis, G. & Charmillot, S. (2017). Schulische Ungleichheit in der Schweiz. Social Change in Switzerland, (8).

Field, A. (2009). *Discovering statistics using SPSS, 3rd edition*. London: Sage.

Friedrich, H. F. & Mandl, H. (1992). Lern- und Denkstrategien – ein Problemaufriss. In H. Mandl & H. F. Friedrich (Hrsg.), *Lern- und Denkstrategien* (*S.* 3–54). Göttingen: Hogrefe.

Gao, X., Shavelson, R. J. & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: promises and problems. *Applied Measurement in Education*, *7*(4).

Geiser, C. (2012). *Data analysis with Mplus.* New York: Guilford.

Germann, P. J. & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching*, *33*(7).

Gick, M. L. & Holyoak, K. J. (1980). Analogical Problem Solving. *Cognitive Psychology, 12*, 306–355.

Glaser, R. (1972). Individuals and learning. The new aptitudes. *Educational Researcher, 1*, 5–13.

Gott, R. & Welford, G. (1987). The assessment of observation in science. *School Science Review*, 69 (247), 217-227.

Gott, R. & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education, 18*(7), 791–806.

Gott, R. & Duggan, S. (2002). Problems with the assessment of performance in practical science: which way now? *Cambridge Journal of Education, 32*(2).

Gut, C. (2012). *Modellierung und Messung experimenteller Kompetenz: Analyse eines large-scale Experimentiertests*, *Bd. 134*. Dissertation. Berlin: Logos.

Gut, C., Metzger, S., Hild, P. & Tardent, J. (2014). Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen. *PhyDiD B.* http://phydid.physik.fu-berlin.de/index.php/phydid-b/article/view/532 (letzter Zugriff am 21.5.2019)

Gut, C., Hild, P., Metzger, S. & Tardent, J. (2017). Vorvalidierung des ExKoNawi-Modells. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 324–331). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Zürich 2016.

Gut, C. & Mayer, J. (2018). Kompetenz. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschafts-didaktischen Forschung* (S. 121–140). Berlin: Springer.

Gwet, K. (2008). Computing inter-rater reliability in the presence of high agreement. *British Journal of Mathematical & Statistical Methodology, 61*(1), 29–48.

Habig, S., van Vorst, H., & Sumfleth, E. (2018). Merkmale kontextualisierter Lernaufgaben und ihre Wirkung auf das situationale Interesse und die Lernleistung von Schüler*innen. *ZfDN*, 24(99)

Haertel, E. H. & Linn, R. L. (1996). Comparability. In G. W. Phillips (ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington D.C.: National Center for Education Statistics.

Hammann, M., Phan, T. T. H. & Bayrhuber H. (2007). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? *Zeitschrift für Erziehungswissenschaft, 10*(8), 33–49.

Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing pupils skills in experimentation. *Journal of Biological Education, 42*(2), 66–72.

Harks, B., Rakoczy, K., Hattie, J., Besser, M., Klieme, E. (2014a). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology: an international journal of experimental educational psychology, 34*(3), 269–290.

Harks, B., Rakoczy, K., Klieme, E., Hattie, J. & Besser, M. (2014b). Indirekte und moderierte Effekte von schriftlicher Rückmeldung auf Leistung und Motivation. In H. Ditton & A. Müller (Hrsg.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 163–194). Münster: Waxmann.

Harmon, M., Smith, T. A., Martin, M. O., Kelley, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez E. J. et al. (1997). *Performance assessment in IEA's third international mathematics and science study.* Chestnut Hill, MA: Boston College.

Hart, C., Mulhall, P., Berry, A., Loughran, J. & Gunstone, R. (2000). What is the purpose of this experiment? Or can students learn something from doing experiments? *Journal of Research in Science Teaching, 37*(7), 655–675.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1).

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* New York: Routledge.

Hattie, J. & Wollenschläger, M. (2014). A conceptualization of feedback. In H. Ditton & A. Müller (Hrsg.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 135–149). Münster: Waxmann.

Heitzmann, A. (2012). Lernaufgaben im naturwissenschaftlich-technischen Unterricht. In S. Keller & U. Bender (Hrsg.), *Aufgabenkulturen* (S. 226–239). Seelze: Kallmeyer.

Hild, P., Metzger, S., Parchmann, I. (2014). Individuelle Förderung experimenteller Kompetenzen mit Lernaufgaben. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht.* Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013.

Hild, P., Gut, C., Metzger, S., Tardent, J. (2015). Typenspezifische Kompetenzprogressionen bei hands-on Testaufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität – Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht*. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014.

Hild, P., Brückmann, M. & Gut, C. (2017). Aussagen zur Konstruktvalidität beim experimentellen Problemtyp effektbasiertes Vergleichen (Projekt ExKoNawi). In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 332–335). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Zürich 2016.

Hild, P., Gut, C., Brückmann, M. (2018a). Validating performance assessments in science. Measures that may help to evaluate students' expertise in doing science. *Research in Science & Technological Education*.

Hild, P., Gut, C., Metzger, S. & Tardent, J. (2018b). Zur Generalisierbarkeit bei Experimentiertests. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht – normative und empirische*

Dimensionen (S. 348–351). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Regensburg 2017.

Hild, P., Metzger, S. & Parchmann, I. (2018c). Beurteilung und Förderung experimenteller Kompetenzen anhand von Aufgaben zum effektbasierten Vergleichen. *ChemKon, 3*, 90–97.

Hild, P., Buff, A., Gut, C. & Parchmann, I. (eingereicht). Adaptives kompetenzbezogenes Feedback beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten. Eine empirische Untersuchung zur Wirksamkeit unterschiedlicher Feedbackformen. Zeitschrift für Didaktik der Naturwissenschaften.

Hodson, D. (2009). *Teaching and learning about science: language, theories, methods, history, traditions and values*. Rotterdam: Sense Publ.

Hofstein, A. (2004). The laboratory in chemistry education: thirty years of experience with developments, implementation, and research. *Chemistry Education Research and Practice,* 5(3), 247–264.

Hofstein, A. & Lunetta, V. N. (2004). The Laboratory in Science Education: Foundations for the Twenty-First Century. *Science Education, 88*(1), 28–54.

Hofstein, A., Dkeidek; I., Katchevitch, D., Nahum, T. L., Kipnis, M., Navon, O., Shore, R., Taitelbaum, D. & Mamlok-Naaman, R. (2019). Research and development of inquiry-type chemistry laboratories in Israel. *Israel Journal of Chemistry*, 59, 514-523.

Hofstetter, D. (2017). *Die schulische Selektion als soziale Praxis: Aushandlungen von Bildungsentscheidungen beim Übergang von der Primarschule in die Sekundarstufe I*. Weinheim: Beltz Juventa.

Hungerford, H. R. & Miles, D. T. (1969). A test to measure observation and comparison skills in science. *Science Education, 53*(1), 61–66.

Ivanov, S. (2011). *Naturwissenschaftliche Kompetenz und fachbezogene Einstellungen.* In U. Vieluf, S. Ivanov & R. Nikolova (Hrsg.), KESS 10/11. Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe. HANSE – Hamburger Schriften zur Qualität im Bildungswesen, (Bd. 10, S. 183–214). Münster: Waxmann.

Jaehnig, W. & Miller, M. L. (2007). Feedback types in programmed instruction: a systematic review. *The Psychological Record, 57*(2), 219.

Jovanovic, J., Solano-Flores, G. & Shavelson, R. J. (1994). Performance-based assessments. *Education and Urban Society, 26,* 352–366.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: issues and practice, 18*(2), 5–17.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogenese Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 135–153.

Kingston, N. & Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educational Measurement:issues and practice, 30*(4), 28–37.

Klieme, E. & Warwas, J (2011). Konzepte der individuellen Förderung. *Zeitschrift für Pädagogik, 57*(6), 805–818.

Kluger, A. N. & DeNisi, A. (1996). Feedback intervention theory. *Feedback, 25*(5).

KMK. Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss (Jahrgangsstufe 10).* https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Chemie.pdf (letzter Zugriff am 17.05.2019)

Koenen, J. (2014). *Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen.* Dissertation Duisburg-Essen.

Koenen, J., Kölbach, E., Emden, M. & Sumfleth, E. (2014). Lösungsbeispiele im Chemieunterricht. Entwicklung und Evaluation verschiedener Formen von Lösungsbeispielen. In B. Ralle, S. Prediger, M. Hammann & M. Rothgangel (Hrsg.), *Lernaufgaben entwickeln, bearbeiten und überprüfen: Ergebnisse und Perspektiven der fachdidaktischen Forschung* (S. 139-148). Münster: Waxmann.

Koenen, J., Emden, M. & Sumfleth, E. (2017). Naturwissenschaftlich-experimentelles Arbeiten: Potenziale des Lernens mit Lösungsbeispielen und Experimentierboxen. *Zeitschrift für Didaktik der Naturwissenschaften, 23*(1), 81–98.

Kölbach, E., Maier-Richter, A. & Sumfleth, E. (2014). Lösungsbeispiele – Eine besondere Form von Lernaufgaben zur Unterstützung individuellen Lernens in den Naturwissenschaften. *ChemKon, 22*(1).

Konsortium HarmoS Naturwissenschaften+ (2010). *Naturwissenschaften. Wissenschaftlicher Kurzbericht und Kompetenzmodell.* https://edudoc.educa.ch/static/web/arbeiten/harmos/harmoS_kurzbericht_-neu.pdf (letzter Zugriff am 23.5.2019)

Krammer, K. (2009). *Individuelle Lernunterstützung in Schülerarbeitsphasen. Eine videobasierte Analyse des Unterstützungsverhaltens von Lehrpersonen im Mathematikunterricht.* Dissertation. Berlin: Waxmann.

Krull, N. & Wolfram, C. (2010). Binnendifferenzierung im Alltag einer Hauptschule. Ein Bericht aus der Praxis. *Pädagogik, 11*(10).

Kuhn, D. & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16(11), 866–870.

Kulhavy, R. W. & Stock, W. A. (1989). Feedback in written instruction: the place of response certitude. *Educational Psychology Review, 1*, 279–308.

Kulhavy, R. W., Stock, W. A., Thornton, N. E., Winston, K. S. & Behrens, J. T. (1990). Response feedback, certitude and learning from text. *The British Jounral of Educational Psychology, 60*(2).

Labudde, P., Niedegger, C., Adamina, M. & Gingins, F. (2012). The development, validation, and implementation of standards in science education: chances and difficulties in the Swiss project HarmoS. In S. Bernholt, K. Neumann & P. Nentwig (eds.), *Making it tangible. Learning outcomes in science education* (pp. 235–259). Münster: Waxmann.

Lenhard, W. & Lenhard, A. (2016). *Berechnung von Effektstärken*. Dettelbach: Psychometrica.

Leuders, T. (2014). Modellierungen mathematischer Kompetenzen – Kriterien für eine Validitätsprüfung aus fachdidaktischer Sicht. *Journal für Mathematik-Didaktik, 35*(1).

Lippmann Kung, R. (2005). Teaching the concepts of measurement: An example of a concept-based laboratory course. *Am. J. Phys., 73*(8), 771–777.

Luthiger, H., Wilhelm, M. & Wespi, C. (2014). Entwicklung von kompetenzorientierten Aufgabensets. *Journal für Leherbildung, 3*.

Luthiger, H., Wilhelm, M., Wespi, C. & Wildhirt, S. (Hrsg.) (2018). *Kompetenzförderung mit Aufgabensets. Theorie – Konzept – Praxis.* Bern: hep Verlag.

Lysakowski, R. S. & Walberg, H. J. (1982). Instructaional effects of cues, participation, and corrective feedback: a quantitative synthesis. *American Educational Research Journal, 19*, 559.

Maier, U., Kleinknecht, M. & Metz, K. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerinnen- und Lehrerbildung, 28*(1), 84–96.

Marschner, J. (2011). *Adaptives Feedback zur Unterstützung des selbstregulierten Lernens durch Experimentieren*. Dissertation. Universität Duisburg-Essen, Fakultät für Bildungswissenschaften.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2005). Academic self-concept, interest, grades, and standadized test scores: reciprocal effects models of causal ordering. *Child Development, 76*(2), 397–416.

Mayer, J., Grube, C. & Möller, A. (2008). Kompetenzmodellierung naturwissenschaftlicher Erkenntnisgewinnung. In U. Harms & A. Sandmann (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik, Vol. 3* (S. 63–79). Innsbruck: Studienverlag.

McComas, W. F., Almazroa, H. & Clough, M. P. (1998). The nature of science in science education: an introduction. *Science & Education, 7*, 511–532.

Mercer, N., Dawes, L., Wegerif, R. & Sams, C. (2004). Reasoning as a scientist: ways of helping children to use language to learn science. *British Educational Research Journal*, 30, 359–377.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational measurement, 3rd edition* (pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (ed.), *Technical issues in large-scale performance assessment.* Washington D.C.: National Center for Education Statistics.

Metzger, S. (2013). Desiderate der naturwissenschaftsdidaktischen Forschung. *Beiträge zur Lehrerinnen- und Lehrerbildung, 31*(1), 42–52.

Metzger, S., Gut, C., Hild, P. & Tardent, J. (2014a). Modelling and assessing experimental competence: an interdisciplinary progress model for hands-on assessments. In C. P. Constantinou, N. Papadouris &

A. Hadjigeorgiou (eds.), *E-Book proceedings of the ESERA 2013 conference: science education research for evidence-based teaching and coherence in learning.*

Metzger, S., Hild, P., Gut, C. & Tardent, J. (2014b). Aufgaben und erste Ergebnisse der hands-on Assessments. In S. Bernhold (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht* (S. 174–176). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013. http://www.gdcp.de/images/tagungsbaende/GDCP_Band34.pdf (letzter Zugriff am 23.5.2019)

Meyer, H. (2009). *Leitfaden Unterrichtsvorbereitung, 4te Auflage*. Berlin: Cornelsen.

Meyer, K. & Carlisle, R. (1996). Children as experimenters. *International Journal of Science Education, 18*(2), 231–48.

Millar, R., Gott, R., Lubben, F. & Duggan, S. (1996). Children's performance of investigative tasks in science: a framework for considering progression. In M. Hughes (ed.), *Progression in learning.* Clevedon, UK: Multilingual Matters LTD.

Miller, M. D. (1998). *Generalizability of performance-based assessments*. Washington DC: Council of the Chief State School Officers.

Miller, M. D. & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367–378.

Muckenfuß, H. (1996). *Lernen im sinnstiftenden Kontext. Entwurf einer zeitgemäßen Didaktik des Physikunterrichts.* Berlin: Cornelsen.

Müller, A. & Ditton, H. (2014). Feedback: Begriff, Formen und Funktionen. In H. Ditton & A. Müller (Hrsg.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 11–28). Münster: Waxmann.

Mushquash, C. & O´Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542–547.

Nakagawa, S. & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Review, 82*, 591–605.

Nehring, A., Schwichow, M. & Gut, C. (2019). Symposium Experimentelle ‚Kompetenz' – ein nützliches Konstrukt? In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe*. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Kiel 2018.

Neumann, K., Kauertz, A., Lau, A., Notarp, H. & Fischer, H. E. (2007). Die Modellierung physikalischer Kompetenz und ihrer Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften 13*, 101–121.

Oelkers, J. (2012). Aufgabenkultur und selbstreguliertes Lernen. In S. Keller & U. Bender (Hrsg.), *Aufgabenkulturen* (S. 81–99). Seelze: Kallmeyer.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences,* 13, 423-451.

Peter, E. (2007). Der Ölfläschchen-Versuch. In R. Duit, H. Gropengießer & L. Stäudel (Hrsg.), *Naturwissenschaftliches Arbeiten, Unterricht und Material 5–10, 2te Auflage* (S. 18–21). Seelze-Velber: Friedrich Verlag.

Perfetto, G. A., Bransford, J. D. & Franks, J. J. (1983). Constraints on access in a problem solving context. *Memory & cognition*, *11*(1), 24–31.

Pfeifer, P. (2002). Erkenntniswege in der Chemie und im Chemieunterricht. In P. Pfeifer, B. Lutz & H. J. Bader (Hrsg.), *Konkrete Fachdidaktik Chemie, 3te Auflage* (S. 90–106). München: Oldenbourg Schulbuchverlag.

Pintrich, P. R. & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33.

Priemer, B., Weß, R., & Ludwig, T. (2019). PCK des Argumentierens im naturwissenschaftlichen Unterricht. In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe*. (S. 596) GDCP Jahrestagung in Kiel 2018.

Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., & van Gog, T. (2017). Effects of performance feedback valence on perceptions of invested mental effort. *Learning and Instruction, 51*, 36–46.

Rakoczy, K., Buff, A. & Lipowsky, F. (2005). Teil I. Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Unterrichtsqualität, Lernverhalten und mathematisches Verständnis*. Frankfurt am Main: DIPF

Rakoczy, K., Klieme, E., Bürgermeister, A. & Harks, B. (2008). The interplay between student evaluation and instruction. *Zeitschrift für Psychologie, 216*(2).

Ramseier, E., Labudde, P. & Adamina, M. (2011). Validierung des Kompetenzmodells HarmoS Naturwissenschaften: Fazite und Defizite. *Zeitschrift für Didaktik der Naturwissenschaften, 17*, 7–33.

Reigosa, C., & Jimenez-Aleixandre, M. P. (2007). Scaffolded problem-solving in the physics and chemistry laboratory: difficulties hindering students' assumption of responsibility. *International Journal of Science Education*, 29, 307–329.

Reusser, K. (2005). Problemorientiertes Lernen – Tiefenstruktur, Gestaltungsformen, Wirkung. *Beiträge zur Lehrerinnen- und Lehrerbildung, 23*(2).

Reusser, K. (2015). Aufgaben – Träger der Lerngelegenheiten und Lernprozesse im kompetenzorientierten Unterricht. *Seminar, 4*, 77–101.

Rincke, K. & Wodzinski, R. (2010). Schülerexperimente: Wege und Wirkungen von Unterstützungsmaßnahmen. In D. Höttecke (Hrsg.), *Entwicklung naturwissenschaftlichen Denkens zwischen Phänomen und Systematik* (S. 242–244). Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Dresden 2009.

Ropohl, M. & Scheuermann, H. (2018). Welche Rückmeldungen wirken am besten? Ergebnisse einer empirischen Untersuchung von Rückmeldeformen beim Planen von Experimenten. *Zeitschrift für Didaktik der Naturwissenschaften*.

Rönnebeck, S., Bernholt, S. & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education*, 52(2), 161–197.

Ruiz-Primo, M. A., Baxter, G. P. & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1).

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: an update. *Journal of Research in Science Teaching, 33*(10).

Rumann, S., Fleischer, J., Stawitz, H., Wirth, J. & Leutner, D. (2010). Vergleiche von Profilen der Naturwissenschafts- und Problemlöse-Aufgaben der PISA 2003-Studie. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 315–328.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science, 18*(2), 119–144.

Sandoval, W. A. & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*(3), 345–372.

Schauble, L., Klopfer, L. E. & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*(9).

Schecker, H. & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 45–66.

Scheuermann, H. (2017). *Entwicklung und Evaluation von Unterstützungsmaßnahmen zur Förderung der Variablenkontrollstrategie beim Planen von Experimenten*. Dissertation. Berlin: Logos.

Schiepe-Tiska, A., Rönnebeck, S., Heitmann, P., Schöps, K., Prenzel, M. & Nagy, G. (2017). Die Veränderung der naturwissenschaftlichen Kompetenz von der 9. zur 10. Klasse bei PISA und den Bildungsstandards unter Berücksichtigung geschlechts- und schulartspezifischer Unterschiede sowie der Zusammensetzung der Schülerschaft. *Zeitschrift für Erziehungswissenschaften, 20*, 151–176.

Schmidkunz, H. (1998). Kochsalz – Lebensmittel oder Chemikalie? *Naturwissenschaften im Unterricht Chemie, 46*, 8–11.

Schreiber, N., Theyßen, H., & Schecker, H. (2009). Experimentelle Kompetenz messen?! *Physik und Didaktik in Schule und Hochschule. PhyDid, 3*(8), 92–101.

Schreiber, N., Theyßen, H. & Schecker, H. (2014). Diagnostik experimenteller Kompetenz: Kann man Realexperimente durch Simulationen ersetzen? *Zeitschrift für Didaktik der Naturwissenschaften, 20*(1), 161–173.

Schreiber, N., Theyßen, H. & Schecker, H. (2016). Process-oriented and product-oriented assessment of experimental skills in physics: a comparison. In N. Papadouris et al. (ed.), *Insights from research in*

science teaching and learning, contributions from science education research (pp. 29–43). Switzerland: Springer-Verlag.

Schreiber, N. & Theyßen, H. (2019). Selbstbeurteilungen zur formativen Diagnostik experimenteller Performanzen – was zeichnet genau urteilende Schülerinnen und Schüler aus? *Zeitschrift für Didaktik der Naturwissenschaften*.

Schulz, A., Wirtz, M. & Starauschek, E. (2012). Das Experiment in den Naturwissenschaften. In W. Rieß, M. Wirtz, B. Barzel & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht* (S. 15–38). Münster: Waxmann.

Schwichow, M. (2015). *Förderung der Variablen-Kontroll-Strategie im Physikunterricht.* Dissertation IPN Kiel.

Schwichow, M., Zimmerman, C., Croker, S. & Härtig, H. (2016). What students learn from hands-on activities: hands-on versus paper-and-pencil. *Journal of Research in Science Teaching*, *53*(7), 980–1002.

Selter, C., Hußmann, S., Hößle, C., Knipping, C., Lengnink, K. & Michaelis, J. (Hrsg.) (2017). *Diagnose und Förderung heterogener Lerngruppen. Theorien, Konzepte und Beispiele aus der MINT-Lehrerbildung* (S.11–18). Münster: Waxmann.

Shavelson, R. J. & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166.

Shavelson, R. J., Baxter, G. P. & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347–362.

Shavelson, R. J. (1992). What we've learned about assessing hands-on science. *Educational Leadership.*

Shavelson, R. J, Gao, X. & Baxter, G. P. (1993). Sampling variability of performance assessments. University of Los Angeles, California: National Center for Research on Evaluation, Standards, and Student Testing. Report 142.

Shavelson, R. J. & Ruiz-Primo, M. A. (1998). On the assessment of science achievement – conceptual underpinnings for the design of performance assessments. University of Los Angeles, California.: National Center for Research on Evaluation, Standards, and Student Testing. Report 491.

Shavelson, R. J., Solano-Flores, G. & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. *Evaluation and Programming Planning*, *21*.

Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, *36*(1), 61–71.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M. et al. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*, *21*(4), 295–314.

SKBF (2018). *Bildungsbericht Schweiz 2018.* Aarau: Schweizerische Koordinationsstelle für Bildungsforschung. https://www.skbf-csre.ch/fileadmin/files/pdfs/bildungsberichte/2018/Bildungsbericht_Schweiz_2018.pdf (letzter Zugriff am 14.6.2019)

Solano-Flores, G. (1994). *A logical model for the development of science performance assessments*. Dissertation. University of Santa Barbara.

Solano-Flores, G. & Shavelson, R. J. (1997). Development of performance assessments in science: conceptual, practical, and logistical issues. *Educational Measurement: issues and practice, 16*(3), 16–24.

Solano-Flores, G., Shavelson, R. J., Schultz, S. E. & Wiley, E. W. (1997). On the development and scoring of classification and observation science performance assessments. University of Los Angeles, California.: National Center for Research on Evaluation, Standards, and Student Testing. Report 458.

Solano-Flores, G., Javanovic, J., Shavelson, R. J. & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293–315.

Sommer, K., Wambach-Laicher, J. & Pfeifer, P. (Hrsg.). (2018). *Konkrete Fachdidaktik Chemie, 3te Auflage* (S. 460–517). München: Oldenbourg.

Stäudel, L., Blum, S., Franke-Braun, G., Hänze, M., Schmidt-Weigand, F. & Wodzinski, R. (2010). *Aufgaben mit gestuften Hilfen für den Chemieunterricht, 2te Auflage*. Seelze: Friedrich Verlag GmbH.

Stecher, B. M. (1996). *Performance assessments in science*. Santa Monica: RAND.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J. & Haertel, E. H. (2000). The effects of content, format, and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13*(2), 139–60.

Stone, C. A. (1998). The metaphor of scaffolding: Its utility for the field of learning disabilities. *Journal of Learning Disabilities*, 31, 344-364.

Strijbos, J. W., Pat-El, R. J., & Narciss, S. (2010). Validation of a (peer) feedback perceptions questionnaire. *Proceedings of the 7th International Conference on Networked Learning*, Aalborg University, Aalborg.

Tabak, I., & Baumgartner, E. (2004). The teacher as partner: exploring participant structures, symmetry, and identity work in scaffolding. *Cognition and Instruction*, 22(4), 393-429.

Taut, S. & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction, 46*, 45–60.

Tesch, M. & Duit, R. (2004). Experimentieren im Physikunterricht – Ergebnisse einer Videostudie. *Zeitschrift für Didaktik der Naturwissenschaften*, 10, 51-69.

Toh, K.-A. & Woolnough, B. E. (1990). Assessing, through reporting, the outcomes of scientific investigations. *Educational Research, 32*(1), 59–65.

Tomera, A. N. (1974). Transfer and retention of transfer of the science process of observation and comparison in Junior high school students. *Science Education, 58*(2).

Tweney, R. D. (1990). Five questions for computationalists. In J. Shrager & P. Langley (eds.), *Computational models of scientific discovery and theory formation*. San Mateo, California: Morgan Kaufmann.

Van de Pol, J., Volman, M. & Beishuizen, J. (2010). Scaffolding in teacher – student interaction: a decade of research. *Educational Psychology Review, 22*(3), 271–96.

Vollmeyer, R. & Rheinberg, F. (2005). A surprising effect of feedback on learning. *Learning and Instruction, 15*, 589–602.

Volkwyn, T. S., Allie, S., Buffler, A. & Lubben, F. (2008). Impact of a conventional introductory laboratory course on the understanding of measurement. *Phys. Rev. ST Physics Educational Research, 4.*

Vorholzer, A. (2016). *Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? Eine empirische Untersuchung eines expliziten und eines impliziten Instruktionsansatzes.* Dissertation. Berlin: Logos.

Vorholzer, A., von Aufschnaiter, C. & Kirschner, S. (2016). Entwicklung und Erprobung eines Tests zur Erfassung des Verständnisses experimenteller Denk- und Arbeitsweisen. *Zeitschrift für Didaktik der Naturwissenschaften*.

Vorholzer, A. & von Aufschnaiter, C. (2019). Guidance in inquiry-based instruction – an attempt to disentangle a manifold construct. *International Journal of Science Education*, 41(11), 1562-1577.

Wahser, I. & Sumfleth, E. (2008). Training experimenteller Arbeitsweisen zur Unterstützung kooperativer Kleingruppenarbeit im Fach Chemie. *Zeitschrift für Didaktik der Naturwissenscahften*, *14*, 219–241.

Walpuski, M. (2006). Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback. Berlin: Logos.

Walpuski, M. & Sumfleth, E. (2007). Strukturierungshilfen und Feedback zur Unterstützung experimenteller Kleingruppenarbeit im Chemieunterricht. *Zeitschrift für Didaktik der Naturwissenscahften*, *13*, 181–198.

Webb, N. M., Schlackman, J. & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13*(3), 277–301.

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In *Handbook of Statistics, Volume 26, 1st edition* (pp. 81–124). Elsevier Psychometrics.

Weinert, F. E. (1999). *Konzepte der Kompetenz.* Paris: OECD.

Wellnitz, N. & Mayer, J. (2008). Evaluation von Kompetenzstruktur und -niveaus zum Beobachten, Vergleichen, Ordnen und Experimentieren. In D. Krüger, A. Upmeier zu Belzen, T. Riemeier & K. Niebert, *Erkenntnisweg Biologiedidaktik* (S. 129–144). Kassel.

Wigfield, A. & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81.

Wilhelm, M. & Kunz, P. (2016). Praktisch-naturwissenschaftliches Arbeiten im Unterricht. In S. Metzger, C. Colberg & P. Kunz (Hrsg.), *Naturwissenschafts-didaktische Perspektiven. Naturwissenschaftliche Grundbildung und didaktische Umsetzung im Rahmen von SWiSE* (Band 1, S. 126–140). Bern: Haupt.

William, D. (2011). The case for formative assessment. In D. William (ed.), *Embedded formative assessment*. Bloomington: Solution Tree Press.

Wirth, J., Künsting, J. & Leutner, D. (2009). The impact of goal specificity and goal type on learning outcome and cognitive load. *Computers in Human Behavior*, 299–305.

Wodzinski, R. & Stäudel, L. (2009). *Aufgaben mit gestuften Hilfen für den Physik-Unterricht*. Seelze: Friedrich.

Wollenschläger, M., Möller, J. & Harms, U. (2012). Ist kompetenzlelles Fremdfeedback überlegen, weil es als effektiver wahrgenommen wird? *Unterrichtswissenschaft*, *40*(3).

Wu, H. K. & Krajcik, J. S. (2006). Exploring middle school students' use of inscriptions in project-based science classrooms. *Science & Education* 90, 852–873.

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., Tomita, M. K. & Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education*, *21*(4), 335–359.

Abbildungsverzeichnis

Tabellenverzeichnis

Publikationen

Zeitschriften- und Buchbeiträge

Hild, P., Buff, A., Gut, C & Parchmann, I. (*im Review-Verfahren*). Adaptives kompetenzbezogenes Feedback beim selbstständigen praktisch-naturwissenschaftlichen Arbeiten. Eine empirische Untersuchung zur Wirksamkeit unterschiedlicher Feedbackformen. *ZfDN.*

Emden, M., Engel, S., Hild, P., Kallinna, K. & Murer, L. (2019a). Glänzend gemacht. *Chemie in unserer Zeit (CiuZ), 53*(1), 66–67.

Emden, M. et al. (2019b). Cola-Käse. *CiuZ, 53*(2), 130–131.

Emden, M. et al. (2019c). Wieso Ballonverkäufer bei der Arbeit keine Zitrusfrüchte essen. *CiuZ, 53*(3), 194–195.

Emden, M. et al. (2019d). Schnell ein Eis. *CiuZ, 53*(4), 266-270.

Emden, M., Kallinna, K., Murer, L. & Hild, P. (2019). Wenn Tintenkiller sauer wären. *CiuZ, 53*(5), 350-351.

Emden, M., Kallinna, K., Murer, L. & Hild, P. (2020). Wer braucht schon Froschschenkel. *CiuZ,* 54(1).

Emden, M., Kallinna, K., Murer, L. & Hild, P. (eingereicht). Bloß keinen Schnaps zum Lebkuchen! *CiuZ.*

Adamina, M. & Hild, P. (2019). Mit Lernaufgaben Kompetenzen fördern. In P. Labudde & S. Metzger (Hrsg.), *Fachdidaktik Naturwissenschaften, 1.–9. Schuljahr*, 3te Auflage. Bern: Haupt Verlag.

Hild, P., Gut, C. & Brückmann, M. (2018). Validating performance assessments in science. Measures that may help to evaluate students' expertise in doing science. *Research in Science & Technological Education.*

Hild, P. & Kallinna, K. (2018). Heavy Metal & Co. *RAAbits Naturwissenschaften, 27*. Dr. Josef Raabe Verlags-GmbH.

Hild, P., Metzger, S. & Parchmann, I. (2018). Beurteilung und Förderung experimenteller Kompetenzen mit Aufgaben zum effektbasierten Vergleichen. *ChemKon, 25*(3), 90–97.

Hild, P. & Kölbach E. (2017). Der Schminkkoffer der Alten Ägypter. *RAAbits Naturwissenschaften, 22.* Dr. Josef Raabe Verlags-GmbH.

Brückmann, M., Kölbach, E., Metzger, S. & Hild, P. (2015). Fachdidaktische Weiterbildungen in den Naturwissenschaften: Ausgangslage und Ziele einer praxisorientierten Professionalisierung. *Beiträge zur Lehrerinnen- und Lehrerbildung,* 246–255.

Hild, P., Kölbach, E. & Metzger, S. (2015). Beobachten lernen – Aufgaben zur Förderung der Beobachtungskompetenz. *Naturwissenschaften im Unterricht Chemie, 149.*

Kölbach, E. & Hild, P. (2015). Ist Tee gleich Tee? – Unterscheidung von Tee und Aufgussgetränken mit Schülerexperimenten. *RAAbits Chemie, 30.* Dr. Josef Raabe Verlags-GmbH.

Hild, P. & Schraner, M. (2014). Galvanisches Versilbern einer Signalpfeife. *Praxis der Naturwissenschaften – Chemie in der Schule, 63*(6).

Konferenzbeiträge

Hild, P., Gut, C., Metzger, S. & Tardent, J. (2018). Zur Generalisierbarkeit bei Experimentiertests. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht – normative und empirische Dimensionen* (S. 348–351). Gesellschaft für Didaktik der Chemie und Physik (GDCP), Jahrestagung in Regensburg 2017.

Brückmann, M., Hild, P., Gut, C. & Metzger, S. (2017). ESPri – Studie zur Erhebung von Präkonzepten zum Thema Energie. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 154–157). GDCP, Jahrestagung in Zürich 2016.

Gut, C., Hild, P., Metzger, S. & Tardent, J. (2017).Vorvalidierung des ExKoNawi-Modells. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 324–331). GDCP, Jahrestagung in Zürich 2016.

Hild, P., Brückmann, M. & Gut, C. (2017). Aussagen zur Konstruktvalidität beim experimentellen Problemtyp effektbasiertes Vergleichen (Projekt ExKoNawi). In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 332–335). GDCP, Jahrestagung in Zürich 2016.

Hild, P., Tardent, J., Gut, C. & Metzger, S. (2015). Typenspezifische Kompetenzprogressionen bei hands-on Testaufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität – Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht.* GGDCP, Jahrestagung in Bremen 2014.

Hild, P., Tardent, J., Gut, C. & Metzger, S. (2015). Typenspezifische Kompetenzprogressionen bei hands-on Testaufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität – Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht.* GDCP, Jahrestagung in Bremen 2014.

Gut, C., Metzger, S., Hild, P. & Tardent, J. (2014). Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen. *PhyDiD B.* Beiträge zur DPG-Frühjahrstagung 2014.

Hild, P., Metzger, S., Parchmann, I. (2014). Individuelle Förderung experimenteller Kompetenzen mit Lernaufgaben. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht.* GDCP, Jahrestagung in München 2013.

Metzger, S., Hild, P., Gut, C. & Tardent, J. (2014a). Projekt ExKoNawi: Aufgaben und erste Ergebnisse der hands-on Assessments. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht.* GDCP, Jahrestagung in München 2013.

Metzger, S., Gut, C., Hild, P. & Tardent, J. (2014b). Modelling and assessing experimental competence: an interdisciplinary progress model for hands-on assessments. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (eds.), *E-Book proceedings of the ESERA 2013 conference: science education research for evidence-based teaching and coherence in learning.*

Erklärung

Hiermit erkläre ich, dass diese Dissertation – abgesehen von der Beratung durch meine Betreuerinnen und Betreuer – nach Inhalt und Form meine eigene Arbeit ist. Alle wörtlichen oder dem Sinne nach entnommenen Stellen anderer Veröffentlichungen wurden von mir gekennzeichnet. Die Dissertation ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der deutschsprachigen *scientific community* entstanden. Die Arbeit hat weder im Ganzen noch in Teilen an anderer Stelle im Rahmen eines Promotionsverfahrens vorgelegen. Ein Teil der Ergebnisse dieser Dissertation (Kapitel 3–6) wurde bereits in Form von Veröffentlichungen publiziert bzw. zur Publikation eingereicht. Dem Autor wurde niemals ein akademischer Grad entzogen.

Zürich, den _____ _____

Pitt Hild