Studien zum Physik- und Chemielernen

M. Hopf und M. Ropohl [Hrsg.]

355

Livia Murer

Diagnose experimenteller Kompetenzen beim praktisch-naturwissenschaftlichen Arbeiten

Vergleich verschiedener Methoden und kognitive Validierung eines Testverfahrens

λογος

Studien zum Physik- und Chemielernen

Herausgegeben von Martin Hopf und Mathias Ropohl

Diese Reihe im Logos Verlag Berlin lädt Forscherinnen und Forscher ein, ihre neuen wissenschaftlichen Studien zum Physik- und Chemielernen im Kontext einer Vielzahl von bereits erschienenen Arbeiten zu quantitativen und qualitativen empirischen Untersuchungen sowie evaluativ begleiteten Konzeptionsentwicklungen zu veröffentlichen. Die in den bisherigen Studien erfassten Themen und Inhalte spiegeln das breite Spektrum der Einflussfaktoren wider, die in den Lehr- und Lernprozessen in Schule und Hochschule wirksam sind.
Die Herausgeber hoffen, mit der Förderung von Publikationen, die sich mit dem Physik- und Chemielernen befassen, einen Beitrag zur weiteren Stabilisierung der physik- und chemiedidaktischen Forschung und zur Verbesserung eines an den Ergebnissen fachdidaktischer Forschung orientierten Unterrichts in den beiden Fächern zu leisten.

Martin Hopf und Mathias Ropohl

Studien zum Physik- und Chemielernen

Band 355

Livia Murer

Diagnose experimenteller Kompetenzen beim praktisch-naturwissenschaftlichen Arbeiten

Vergleich verschiedener Methoden und kognitive Validierung eines Testverfahrens

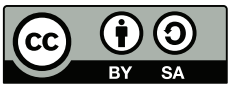
Logos Verlag Berlin

Studien zum Physik- und Chemielernen
Martin Hopf und Mathias Ropohl [Hrsg.]

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Publiziert mit Unterstützung des Schweizerischen Nationalfonds zur Förderung der wissenschaftlichen Forschung.



ISBN 978-3-8325-5657-0
DOI 10.30819/5657
ISSN 1614-8967

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10
D-12681 Berlin
Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
https://www.logos-verlag.de

Danksagung

An dieser Stelle möchte ich mich bei all den Menschen bedanken, die mich während meiner Promotion begleitet und unterstützt haben. Ich bin für sehr vieles dankbar und eine Seite reicht nicht aus, um alles aufzuzählen. Darum möchte ich in der Folge einige aus meiner Sicht besonders wichtige Punkte hervorheben.

Allen Voran möchte ich mich bei Susanne Metzger bedanken. Danke für die fortwährende Unterstützung, die kritischen und immer konstruktiven Hinweise, dafür, dass Du Dir immer Zeit für mich genommen hast und dass auf Dich immer Verlass ist und für die freundschaftlichen und persönlichen Gespräche. Du hast massgebend dazu beigetragen, dass ich die Promotionszeit positiv erlebt habe.

Besonderen Dank gilt auch Andreas Vorholzer. Danke für die spannenden Diskussionen, die klugen Hinweise, die vielen helfenden Kommentare und dafür, dass auch Du Dir immer Zeit für einen inhaltlichen und methodischen Austausch genommen hast.

Ein grosser Dank gilt auch meinen Kollegen der Didaktik der Naturwissenschaften, Markus Emden und Pitt Hild. Bei Fragen konnte ich stets auf Euch zukommen und Ihr habt mich mit hilfreichen Hinweisen unterstützt.

Mein Dank gilt auch den Kolleginnen der Gesamtvalidierungsstudie des Projekts ExKoNawi, Angela Bonetti und Kirsten Kallinna. Gemeinsam konnten wir die Erhebung und Kodierung der Daten realisieren und Ihr habt mich stets in meinem Vorhaben unterstützt.

Mein Dank gilt den Schulen, die sich dazu bereit erklärt haben, an der Studie teilzunehmen. Ich danke vor allem den Schülerinnen und Schülern, die bereitwillig und engagiert experimentiert und bei der Video- und Interviewstudie teilgenommen haben.

Ganz besonders danke ich meiner Familie. Ihr musstet in vielen Stunden auf mich zu Gunsten meiner Promotion verzichten. Allen Voran möchte ich meinem Mann Roman und meinen Kindern Linus und Laurin danken: Ihr habt mich mit viel Verständnis und Geduld während dieser Zeit unterstützt. Besonderen Dank geht auch an meine Eltern und Schwiegereltern, die uns bei der Betreuung der Kinder geholfen und mich so wesentlich bei der Promotion unterstützt haben.

Inhaltsverzeichnis

Empirischer Teil

1. Einleitung

Das gesellschaftliche Leben ist zunehmend durch Naturwissenschaften geprägt. Naturwissenschaftliche Bildung ist somit von zentraler Bedeutung, um an der Gesellschaft teilzuhaben (vgl. z. B. National Research Council (NRC), 2012; Schiepe-Tiska et al., 2016; Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK), 2005). Es besteht weitgehend Einigkeit, dass naturwissenschaftliche Bildung nicht nur Kenntnisse inhaltsbezogener Kompetenzen umfasst, sondern dass auch naturwissenschaftliche Denk- und Arbeitsweisen von zentraler Bedeutung sind, was sich in nationalen und internationalen Bildungsstandards und Lehrplänen widerspiegelt (z. B. Deutschschweizer Erziehungsdirektorenkonferenz (D-EDK), 2016; Erziehungsdirektorenkonferenz (EDK), 2011; NRC, 2012; KMK, 2005). Bei den naturwissenschaftlichen Denk- und Arbeitsweisen sollen die Schülerinnen und Schüler unter anderem experimentelle Kompetenzen aufbauen. Dabei stellt sich die Frage, was experimentelle Kompetenzen genau sind respektive welche Teilkompetenzen diese umfassen und wie diese diagnostiziert werden können, damit ein gezieltes Aufbauen und Fördern ermöglicht wird. Der Fokus vorliegender Arbeit liegt auf der Diagnose experimenteller Kompetenzen.

Welche Teilkompetenzen experimenteller Kompetenzen unterschieden werden, ist in Studien sehr heterogen (z. B. Emden, 2011; Maiseyenka, 2014; Metzger et al., 2014; Schreiber, 2012; Wellnitz & Mayer, 2013). Da die vorliegende Arbeit in das vom Schweizerischen Nationalfond finanzierte Projekt ExKoNawi (*Ex*perimentelle *Ko*mpetenzen in den *Na*tur*wi*ssenschaften) eingebettet ist, orientiert sich auch die Definition der Teilkompetenzen experimenteller Kompetenzen am Projekt ExKoNawi: Die Teilkompetenzen beruhen auf den Fähigkeiten, unterschiedliche experimentelle Problemstellungen lösen zu können. Entsprechend werden sie auch Problemtypen genannt. Dabei wird im Rahmen des Projekts zwischen den Problemtypen «Naturwissenschaftliches Messen mit vorgegebenen Instrumenten» (abgekürzt: «Messen»), «Vergleichende Untersuchung von Objekten» (abgekürzt: «Vergleichen») und «Experimentelles Untersuchen der Beziehungen zwischen Variablen» (abgekürzt: «Untersuchen») differenziert (vgl. Bonetti et al., 2017). Im Rahmen der Arbeit liegt der Fokus auf dem Problemtyp «Messen».

Für einen gezielten Aufbau experimenteller Kompetenzen müssen diese möglichst genau diagnostiziert werden. Zur Diagnose experimenteller Kompetenzen existieren verschiedene Testarten, zum Beispiel schriftliche Tests, Tests mit Computersimulationen oder Tests mit Realexperimenten. Die Diagnose

experimenteller Kompetenzen erfolgt in vielen Studien, unter anderem aus testökonomischen Gründen, mithilfe schriftlicher Testverfahren (z. B. Hammann et al., 2008; Mannel, 2011; Mayer et al., 2008). Allerdings zeigen Studien, dass mit rein schriftlichen Testverfahren einige handlungsbezogene experimentelle Kompetenzen (wie z. B. das Handhaben von Messinstrumenten) nur unzureichend abgebildet werden können (vgl. z. B. Emden, 2011; Hammann et al., 2008; Schreiber, 2012; Shavelson et al., 1999). Unter anderem deshalb werden Tests mit Realexperimenten als Benchmark zur Diagnose experimenteller Kompetenzen betrachtet (vgl. z. B. Baxter & Shavelson, 1994; Wenning, 2007). Somit braucht es für die Forschung und den naturwissenschaftlichen Unterricht valide Testverfahren mit Realexperimenten. Ein Ziel vorliegender Arbeit besteht dementsprechend darin, ein Testverfahren mit Realexperimenten am Beispiel der Aufgaben des Problemtyps «Messen» zu validieren, indem untersucht wird, inwiefern die Schülerinnen und Schüler zum Lösen der Aufgaben hauptsächlich die intendierten Konzepte verwenden und die Aufgaben somit kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen.

Bei Tests mit Realexperimenten können die experimentellen Kompetenzen durch verschiedene Erhebungsmethoden diagnostiziert werden, zum Beispiel durch Schülerprotokolle[1], Beobachtungen während des Experimentierens (z. B. auch Videos) oder die Rekonstruktion von Denkprozessen von Schülerinnen und Schülern (z. B. aus Interviews oder durch die Methode des Lauten Denkens). Während Beobachtungen und die Rekonstruktion von Denkprozessen von Schülerinnen und Schülern sehr zeitintensiv sind und sich auf Klassenebene aufgrund der Gruppengrösse nur schwierig realisieren lassen (vgl. auch Emden & Sumfleth, 2012), stellen Schülerprotokolle eine ökonomische Alternative dar. Schülerprotokolle geben aber nur eine indirekte Aussage über handlungsbezogene experimentelle Kompetenzen (vgl. z. B. Abrahams et al., 2013) und es muss sichergestellt werden, dass die Schülerinnen und Schüler einerseits protokollieren, was sie gemacht haben und anderseits gemacht haben, was sie protokollieren (Gut-Glanzmann, 2012). Somit stellt sich die Frage, durch welche Erhebungsmethoden die experimentellen Kompetenzen bei Tests mit Realexperimenten möglichst genau, aber dennoch ökonomisch, erfasst werden können. Ein weiteres Ziel der Arbeit besteht demzufolge darin zu untersuchen, inwiefern experimentelle Kompetenzen mit Schülerprotokollen aus Tests mit Realexperimenten genau

[1] In vorliegender Arbeit werden Schülerinnen- und Schülerprotokolle wegen der besseren Lesbarkeit als Schülerprotokolle bezeichnet.

diagnostiziert werden können und in welcher Hinsicht zusätzliche Erhebungsmethoden die Genauigkeit der Diagnostik erhöhen.

Fragestellungen zur Diagnose experimenteller Kompetenzen können nicht losgelöst von Fragen zur Definition und Modellierung experimenteller Kompetenzen untersucht werden. Somit wird in vorliegender Arbeit zunächst auf den Stand der Forschung zur Definition und Modellierung experimenteller Kompetenzen eingegangen (Kapitel 2). Dabei werden verschiedene Möglichkeiten beschrieben, experimentelle Kompetenzen anhand verschiedener Teilkompetenzen zu modellieren, und anhand exemplarischer Studien illustriert. Ausserdem wird auf das naturwissenschaftliche Messen als Teilkompetenz experimenteller Kompetenzen eingegangen, da im Rahmen vorliegender Arbeit der Fokus auf dem Problemtyp «Messen» liegt. Anschliessend wird der Forschungsstand zur Diagnose experimenteller Kompetenzen erläutert (Kapitel 3), indem verschiedene Testarten und Erhebungsmethoden zur Erfassung experimenteller Kompetenzen beschrieben und bezüglich ihrer Charakteristika verglichen werden. Die theoretischen Grundlagen abschliessend wird definiert, was unter Validität beziehungsweise den verschiedenen Validitätsaspekten nach Messick (1995) zu verstehen ist (Kapitel 4). Hierbei wird insbesondere auf die kognitive Validität fokussiert, da im Rahmen vorliegender Arbeit ein Testverfahren am Beispiel der Aufgaben des Problemtyps «Messen» kognitiv validiert wird. Ausgehend von den theoretischen Grundlagen und dem Stand der Forschung werden dann in Kapitel 5 die Forschungsfragen und Hypothesen hergeleitet. Der darauffolgende empirische Teil der Arbeit gliedert sich in drei Teile. Zunächst werden die Stichprobe, der Ablauf der Datenerhebung und die Instrumente beschrieben, die Teilstudie I (zum Vergleich verschiedener Erhebungsmethoden bei Tests mit Realexperimenten) und Teilstudie II (zur kognitiven Validierung eines Testverfahrens) gemein sind (Kapitel 6). Daraufhin werden die Vorgehensweisen zur Auswertung der Daten und die Ergebnisse im Rahmen von Teilstudie I (Kapitel 7) und Teilstudie II (Kapitel 8) dargestellt. In einem abschliessenden Kapitel werden schliesslich die Ergebnisse beider Teilstudien zusammengefasst und Implikationen für den naturwissenschaftlichen Unterricht und die fachdidaktische Forschung abgeleitet (Kapitel 9).

Allgemeine theoretische Grundlagen der Arbeit

2. Experimentieren als naturwissenschaftliche Methode der Erkenntnisgewinnung und experimentelle Kompetenzen

In den folgenden Ausführungen wird zuerst auf das Experimentieren als naturwissenschaftliche Methode der Erkenntnisgewinnung eingegangen (Unterkapitel 2.1). Ausgehend von einem Begriffsverständnis von Experimentieren können Kompetenzerwartungen definiert und experimentelle Kompetenzen anhand verschiedener Teilkompetenzen modelliert werden. Verschiedene Möglichkeiten zur Modellierung experimenteller Kompetenzen werden in Unterkapitel 2.2 vorgestellt. Das naturwissenschaftliche Messen als Teilkompetenz experimenteller Kompetenzen wird daraufhin in Unterkapitel 2.3 erläutert, da im Rahmen vorliegender Arbeit der Fokus auf dem Problemtyp «Messen» liegt.

2.1 Experimentieren als naturwissenschaftliche Methode der Erkenntnisgewinnung

In den Naturwissenschaften liegen verschiedene Definitionen für das Experimentieren vor (vgl. z. B. Barzel et al., 2012; Höttecke & Rieß, 2015; Metzger et al., 2019; Wellnitz & Mayer, 2013; Wilhelm & Kunz, 2016). So schreiben beispielsweise Höttecke und Rieß (2015): «Die Wissenschaftsforschung hat auf die Frage, was wir unter einem Experiment in den Naturwissenschaften verstehen können, keine eindeutige, sondern viele Antworten» (S. 136). Allgemein wird unter Experimentieren ein objektives und wiederholbares Verfahren der Erkenntnisgewinnung verstanden (Gut & Mayer, 2018). In einem breit aufgefassten Begriffsverständnis von Experimentieren zählen alle naturwissenschaftlichen Methoden dazu, bei welchen eine handelnde Auseinandersetzung mit der Natur stattfindet, um Daten zu gewinnen und diese vor dem Hintergrund von Theorien zu interpretieren und dadurch neue Erkenntnisse über die Natur abzuleiten (Gut & Mayer, 2018). Bei diesem Begriffsverständnis von Experimentieren zählen auch Methoden wie das Beobachten oder Vergleichen zu den experimentellen Zugängen. Wird das Begriffsverständnis von Experimentieren enger gefasst, dann wird beim Experimentieren ein systematisches Eingreifen in das Geschehen vorausgesetzt. Somit unterscheidet sich die experimentelle Vorgehensweise von anderen naturwissenschaftlichen Methoden, wie beispielsweise dem Beobachten (Schulz et al., 2012). Wesentliche Merkmale des Experimentierens sind beim enger gefassten Begriffsverständnis das Untersuchen von kausalen Zusammenhängen (Ursache-Wirkungs-Beziehungen) und die Variablenkontrollstrategie. Bei der

Variablenkontrollstrategie wird die potenzielle Ursache aktiv verändert, um daran den potenziellen Effekt zu erkennen. Alle anderen potenziellen alternativen Einflussfaktoren werden kontrolliert, das heisst konstant gehalten (Schulz et al., 2012).

Gegenwärtig zeigt sich, dass das Begriffsverständnis von Experimentieren sehr divers ist und in den verschiedenen Disziplinen der Naturwissenschaftsdidaktik der Begriff oft auch unterschiedlich aufgefasst wird (Metzger et al., 2019). Während in der Biologiedidaktik meistens von einem engen Begriffsverständnis ausgegangen wird, mit dem Untersuchen von kausalen Zusammenhängen und dem Verwenden der Variablenkontrollstrategie (z. B. Wellnitz & Mayer, 2012), wird das Experimentieren in der Chemie- oder Physikdidaktik scheinbar breiter aufgefasst. So spielen bei den experimentellen Zugängen beispielsweise in der Chemiedidaktik verfahrensbasierte Tests (z. B. Emden, 2011) und in der Physikdidaktik das Herstellen von Messvorrichtungen (z. B. Schreiber, 2012) eine zentrale Rolle (Gut & Mayer, 2018).

Im Rahmen der Arbeit, wie auch beim Projekt ExKoNawi, werden experimentelle Kompetenzen *interdisziplinär* aufgefasst (vgl. z. B. Gut, Metzger, et al., 2014). Entsprechend wird in der vorliegenden Arbeit ebenfalls von einem breiten Begriffsverständnis von Experimentieren ausgegangen, wobei eine handelnde Auseinandersetzung mit der Natur vorausgesetzt wird, um Daten zu gewinnen und diese vor dem Hintergrund von Theorien zu interpretieren. Dabei zählen auch Vorgehensweisen wie das kriteriengeleitete Vergleichen oder das naturwissenschaftliche Messen zu den experimentellen Zugängen, wobei im Rahmen dieser Arbeit der Fokus auf dem Messen liegt.

2.2 Modellierung experimenteller Kompetenzen

Ausgehend von einem Begriffsverständnis von Experimentieren können Kompetenzerwartungen, die auch experimentelle Kompetenzen genannt werden, definiert werden. Die experimentellen Kompetenzen können dabei in weitere Teilkompetenzen zerlegt werden. Dabei gibt es verschiedene Möglichkeiten, die Teilkompetenzen zu definieren und somit die experimentellen Kompetenzen zu modellieren (Gut & Mayer, 2018). Allgemein kann bei der Modellierung von Kompetenzen zwischen Kompetenzstruktur- und Kompetenzentwicklungsmodellen differenziert werden (vgl. z. B. Schecker & Parchmann, 2006; von Aufschnaiter & Rogge, 2010). Kompetenzstrukturmodelle stellen die zu erreichenden Kompetenzen geordnet dar. Im Vergleich dazu machen Kompetenzentwicklungsmodelle auch Annahmen darüber, in welcher Weise sich Kompetenzstrukturen herausbilden, womit die Entwicklung der Kompetenzen genauer in den

Blick genommen wird (Schecker & Parchmann, 2006). Des Weiteren kann zwischen normativ und deskriptiv entwickelten Modellen unterschieden werden. Bei normativen Modellen werden die Komponenten des Modells theoretisch entwickelt und normativ begründet. Im Vergleich dazu wird ein deskriptives Modell aus empirischen Ergebnissen abgeleitet (vgl. z. B. Schaper, 2009; Schecker & Parchmann, 2006). Bei der Modellierung experimenteller Kompetenzen werden zudem im Wesentlichen zwei Ansätze diskutiert, um Teilkompetenzen abzugrenzen. Diese zwei Ansätze werden auch als Teilprozess- und Problemtypenansatz bezeichnet (vgl. z. B. Gut, Hild, et al., 2014; Gut & Mayer, 2018).

Beim Teilprozessansatz (z. B. Emden & Sumfleth, 2012; Vorholzer et al., 2020) werden experimentelle Teilkompetenzen als experimentelle Teilprozesse aufgefasst. Modelle, die vom Teilprozessansatz ausgehen, basieren häufig auf einem Modell von Klahr (Klahr, 2000; Klahr & Dunbar, 1988). Klahr (2000) beschreibt beim Modell «Scientific Discovery as Dual Search (SDDS)» den naturwissenschaftlichen Erkenntnisprozess mit Hilfe von zwei Suchräumen, dem Hypothesen- und Experimentiersuchraum. Im Hypothesensuchraum wird eine Hypothese gesucht und gewählt. Im Experimentiersuchraum wird ein Experiment gesucht, das interpretierbare Ergebnisse liefert, um die Hypothese zu überprüfen. Im letzten Schritt im Modell von Klahr (2000) werden die erhaltenen Ergebnisse interpretiert und auf die Eingangshypothese bezogen. So wird gegebenenfalls über weitere Suchen im Hypothesen- oder Experimentiersuchraum entschieden. Im Modell von Klahr (2000) wird somit zwischen drei Teilprozessen differenziert: Hypothesenbildung, Planung von Experimenten und Schlussfolgerung aus experimentellen Daten. Auch wenn die Modelle mit Teilprozessansatz häufig auf dem Modell von Klahr (2000) basieren, gibt es dennoch sehr unterschiedliche Modellierungen, die sich in der Anzahl und Benennung der Teilprozesse unterscheiden. Übersichten über verschiedene Modellierungen innerhalb des Teilprozessansatzes geben beispielsweise Emden (2011, S. 11) und Schreiber (2012, S. 28). In einigen Modellen werden so beispielsweise drei Teilprozesse differenziert (z. B. Emden & Sumfleth, 2012; Schreiber, 2012; Vorholzer et al., 2020). Diese werden bei Emden und Sumfleth (2012) als Ideen finden, Experiment durchführen und Schlussfolgern bezeichnet und Vorholzer und andere (2020) unterscheiden zwischen Formulieren von Fragen und Hypothesen, Planen von Untersuchungen und Auswerten und Interpretieren von Daten. Im Vergleich dazu wurden in anderen Studien Modelle mit mehr Teilprozessen entwickelt: Gut-Glanzmann (2012)

differenziert zwischen fünf Teilprozessen[2] und Maiseyenka (2014) unterscheidet sechs experimentelle Teilprozesse[3]. Empirische Untersuchungen zu Modellierungen innerhalb des Teilprozessansatzes liefern zudem ein uneindeutiges Bild (Vorholzer et al., 2016). Während sich in manchen Studien die Teilprozesse als differenzierbare Dimensionen trennen lassen (z. B. Grube, 2011; Klos, 2008), konnten in anderen Studien die definierten Teilprozesse nicht als klar trennbare Dimensionen erfasst werden (z. B. Gut-Glanzmann, 2012; Hammann et al., 2008; Henke, 2007; Wellnitz & Mayer, 2013), was für ein eindimensionales Modell einer umfassenden experimentellen Kompetenz spricht.

Bei der Modellierung experimenteller Kompetenzen mithilfe des Problemtypenansatzes (vgl. z. B. Gut, Metzger, et al., 2014) werden experimentelle Teilkompetenzen als die Fähigkeit aufgefasst, unterschiedliche experimentelle Problemstellungen, wie beispieleisweise eine naturwissenschaftliche Messung oder einen kriteriengeleiteter Vergleich, lösen zu können. Diese experimentellen Problemstellungen können auch als Problem- oder Aufgabentypen bezeichnet werden und die experimentellen Kompetenzen werden bei diesem Ansatz innerhalb eines Problemtyps modelliert (Gut, Metzger, et al., 2014). Innerhalb eines Problemtyps wird gemeinsames Strategiewissen zur Problemlösung gebraucht, beispielsweise bei Problemstellungen zum naturwissenschaftlichen Messen wird Strategiewissen im Bereich des Durchführens von Messwiederholungen oder der Interpretation von Messdaten benötigt (vgl. z. B. Gott & Dugan, 2002). Dieses Strategiewissen kann innerhalb eines Problemtyps von einer Problemstellung auf eine andere transferiert werden, indem das durch das Lösen einer Problemstellung erworbene Strategiewissen beim Lösen einer anderen Problemstellung des gleichen Problemtyps angewendet werden kann. Einige Studien (z. B. Gott & Dugan, 1995; Nehring et al., 2014; Ruiz-Primo & Shavelson, 1996; Wellnitz & Mayer, 2013) unterscheiden zwischen verschiedenen experimentellen Problemstellungen. Gott und Dugan (1995) unterscheiden beispielsweise sechs verschiedene Problemstellungen: (1) Variablenbasiertes Untersuchen (den Einfluss einer oder mehrerer unabhängiger Variablen auf eine abhängige Variable untersuchen); (2) Logisches Schlussfolgern (Problemstellungen, deren Lösung verschiedene Teilschritte erfordern); (3) Messen (eine vorgegebene Variable quantitativ messen); (4) Konstruktionen (Ingenieuraufgaben, finden einer Problemlösung mit anschliessender Überprüfung ihrer Effektivität); (5) Konstruktionen

[2] 5 Teilprozesse: Fragen und Hypothesen formulieren; Untersuchungen planen; Untersuchungen durchführen; Daten auswerten und Untersuchungen reflektieren (vgl. Gut-Glanzmann, 2012).

[3] 6 Teilprozesse: Fragestellung entwickeln; Hypothesen bilden; Experiment planen; Beobachten / Messen / Dokumentieren; Daten aufbereiten und Schlüsse ziehen (vgl. Maiseyenka, 2014).

(technologische Aufgaben, eine Konstruktion erstellen, die den gegebenen Anforderungen genügt) und (6) Explorationen (offene Aufgaben, bei denen selbstständig Fragestellungen und Ziele formuliert werden) (Gott & Dugan, 1995; übersetzt von Arndt, 2016). Auch Ruiz-Primo und Shavelson (1996) haben festgestellt, dass Aufgaben mit einem Realexperiment aufgrund ihrer zugrundeliegenden Problemstellung in unterschiedliche Typen eingeteilt werden können: «We have discovered […] a small number of different types of tasks that characterize a wide variety of science performance assessments […]: (a) comparative – compare two or more objects on some attribute while controlling other variables […]; (b) component identification – determine the components that make up the whole […]; (c) classification – create a classification scheme using attributes of a set of objects and a particular goal for classification […]; (d) observation – observe and systematically record an attribute of an object over a period of time […]. […] Additional research may very well expand our category system or a new system may replace it.» (Ruiz-Primo & Shavelson, 1996, S. 1053). Ruiz-Primo und Shavelson (1996) unterscheiden somit vier unterschiedliche Problemstellungen (Vergleichende Untersuchung, Identifizieren von Komponenten eines Ganzen, Klassifizieren und Beobachten) und weisen darauf hin, dass weiterführende Forschung dazu führen kann, dass ihr Ansatz um weitere differenzierbare Problemstellungen ergänzt wird.

In manchen Studien (z. B. Nehring et al., 2014; Wellnitz & Mayer, 2013) werden Teilprozess- und Problemtypenansatz kombiniert, indem sowohl zwischen verschiedenen Problemstellungen als auch Teilprozessen differenziert wird. Nehring und andere (2014) differenzieren beispielsweise in ihrem Modell die naturwissenschaftlichen Arbeitsweisen Beobachten / Vergleichen / Ordnen, Experimentieren[4] und Modelle nutzen und differenzieren dabei die Teilprozesse Fragestellung und Hypothese, Planung und Durchführung sowie Auswertung und Reflexion. Im Vergleich dazu unterscheiden Wellnitz und Mayer (2013) in ihrem Modell die naturwissenschaftlichen Erkenntnismethoden Experimentieren[5], Beobachten und Vergleichen und die Teilprozesse Fragstellung, Hypothese, Untersuchungsdesign und Datenauswertung. Im Rahmen ihrer Studie haben Wellnitz und Mayer (2013) anhand empirischer Daten die Passung verschiedener Modelle geprüft: (1) ein eindimensionales Modell «Naturwissenschaftliche Untersuchungen», (2) ein dreidimensionales Modell «Naturwissenschaftliche

[4] Bei Nehring und anderen (2014) wird ein enges Begriffsverständnis von Experimentieren vertreten (vgl. Unterkapitel 2.1), mit dem Untersuchen von kausalen Zusammenhängen und der Variablenkontrollstrategie.

[5] Bei Wellnitz und Mayer (2013) wird ebenfalls ein enges Begriffsverständnis von Experimentieren vertreten, vgl. Fussnote 4.

Erkenntnismethoden» (Beobachten, Vergleichen, Experimentieren; vgl. Problemtypenansatz) und (3) ein vierdimensionales Modell «Naturwissenschaftliche Erkenntnisschritte» (Fragstellungen, Hypothese, Untersuchungsdesign, Datenauswertung; vgl. Teilprozessansatz). Dabei konnten sie feststellen, dass alle Modelle gleichwertig zu den empirischen Daten passen und bevorzugten das dreidimensionale Modell «Naturwissenschaftliche Erkenntnismethoden», da dieses den aktuellen Forschungsstand am besten miteinbezieht.

Der vorliegenden Arbeit, sowie dem Projekt ExKoNawi, liegt der Problemtypenansatz zugrunde, das bedeutet, dass verschiedene experimentelle Problemstellungen unterschieden werden. Auf die Modellierung experimenteller Kompetenzen im Rahmen vorliegender Arbeit wird in der Folge in Unterkapitel 2.2.1 eingegangen.

2.2.1 Modellierung experimenteller Kompetenzen im Rahmen vorliegender Arbeit

Im Rahmen dieser Arbeit wird wie beim Projekt ExKoNawi von einem breiten Begriffsverständnis von Experimentieren ausgegangen (vgl. Unterkapitel 2.1). Zu den experimentellen Vorgehensweisen zählen somit auch Erkenntnismethoden wie beispielsweise das naturwissenschaftliche Messen oder der kriteriengeleitete Vergleich. Experimentelle Teilkompetenzen werden dabei als die Fähigkeit aufgefasst, unterschiedliche experimentelle Problemstellungen, wie zum Beispiel eine Messung oder einen Vergleich, lösen zu können. Diese unterschiedlichen Problemstellungen werden als Problemtypen bezeichnet. Bei der Modellierung experimenteller Kompetenzen wurde somit von einem interdisziplinären Kompetenzstrukturmodell mit unterschiedlichen Problemtypen ausgegangen (vgl. Metzger & Gut, 2017). Das Modell wurde normativ auf der Basis des Lehrplans und anhand der Analyse bestehender Testinstrumente mit Realexperimenten (z. B. Gut-Glanzmann, 2012 oder Shavelson & Ruiz-Primo, 1999) entwickelt (vgl. Metzger & Gut, 2017). Die Problemtypen wurden in verschiedenen Validierungsstudien mit unterschiedlichen Stichproben geprüft (z. B. Gut et al., 2017; Gut, Metzger, et al., 2014; Hild et al., 2017). Aufgrund dieser Pilotstudien und deren Ergebnisse werden in vorliegender Studie drei Problemtypen unterschieden: «Untersuchen», «Vergleichen» und «Messen» (vgl. auch Bonetti et al., 2017). Diese werden in Tabelle 1 aufgeführt.

Tabelle 1: Modellierung experimenteller Kompetenzen: Beschreibungen der Problemtypen (vgl. Gut & Mayer, 2018, S. 132)

Problemtyp	Problemstellung	Zur Lösung erfordertes Problemverständnis	Zur Lösung erfordertes Strategiewissen	Art der Erkenntnis
«Messen»	Quantitative Grössen mit gegebenen Messinstrumenten möglichst genau messen	Prinzipielle Unsicherheit bei Messungen aufgrund zufälliger Messabweichungen	Messwiederholung und Mittelwertbildung als Strategie, um Messunsicherheiten zu verringern	Quantitative Bestimmung von Merkmalsausprägungen
«Vergleichen»	Objekte anhand einer gegebenen Eigenschaft experimentell vergleichen (ohne direkte Messung der Eigenschaft)	Fairer Vergleich als «Gleiches mit Gleichem auf immer die gleiche Weise vergleichen»	Standardisierung von Vergleichsprozeduren als Strategie, um faire Vergleiche zu ermöglichen	Qualitative Rangfolge von Merkmalsausprägungen und quantifizierbare Reihenfolge
«Untersuchen»	Zusammenhänge zwischen gegebenen Variablen untersuchen	Zusammenhänge als Abhängigkeit einer abhängigen Variablen von unabhängigen Variablen	Variablenkontrolle als Strategie, um einfache Abhängigkeiten zu untersuchen	Korrelative Zusammenhänge zwischen gegebenen Variablen

Nach der Definition der Problemtypen wurden im Rahmen des Projekts ExKoNawi a priori für jeden Problemtypen Qualitätsstandards festgelegt. Diese beschreiben, welche Fähigkeiten und welches Strategiewissen besonders im Fokus stehen (Gut, Metzger, et al., 2014). Im Rahmen vorliegender Arbeit wird auf den Problemtyp «Messen» fokussiert. Somit wird die Modellierung innerhalb eines Problemtyps anhand verschiedener Qualitätsstandards am Beispiel des Problemtyps «Messen» in Unterkapitel 2.3.2 vorgestellt.

2.3 Naturwissenschaftliches Messen als Teilkompetenz experimenteller Kompetenzen

Die genaue Beschreibung von Kompetenzen erfordert, dass man die als zugehörig angenommenen Konzepte benennt (z. B. Vorholzer & von Aufschnaiter, 2020). Diese Konzepte sind Teil des Strategiewissens, welches innerhalb eines Problemtyps von einer Problemstellung auf eine andere Problemstellung des

gleichen Problemtyps transferiert werden kann (vgl. Unterkapitel 2.2). Deshalb wird im Folgenden auf einige zentrale Konzepte im Bereich des naturwissenschaftlichen Messens eingegangen und zudem werden mögliche Schülervorstellungen beziehungsweise Schwierigkeiten der Lernenden erläutert, da diese helfen, die Herangehensweisen und Lösungen der Lernenden beim naturwissenschaftlichen Messen, zum Beispiel auch im Rahmen vorliegender Arbeit, besser zu verstehen.

2.3.1 Zentrale Konzepte im Bereich des naturwissenschaftlichen Messens

Jede Messung wird von einer Unsicherheit begleitet
(Konzept A, vgl. Verweise in Unterkapitel 2.3.2; *vgl. z .B. Fairbrother & Hackling, 1997; Gott et al., 2003; Heinicke, 2012; Hellwig, 2012; Schulz, 2021)*

Ein zentrales Konzept beim naturwissenschaftlichen Messen ist, dass jede Messung von einer Messunsicherheit begleitet wird. So schreiben zum Beispiel Millar und Osborn (1998): «that no observation or measurement can ever be sure of matching exactly the 'true' value – that there is always some uncertainty; that repeating measurements and taking an average is a good method for reducing the effect of random error» (S. 21). Aufgrund der Existenz von Messunsicherheiten ist es in den meisten Fällen nicht möglich, den 'wahren' Wert einer Grösse in Erfahrung zu bringen, sondern man kann sich diesem nur durch geeignete Strategien (z. B. Messwiederholungen und Mittelwertbildung) annähern.

Vor allem jüngere Lernende glauben oft an die Existenz eines 'wahren' Werts: «The idea students have of measurement is associated with the idea of searching for the ‚right' value to be compared to a standard value known by the teacher. For students, any differences between these two values are due to errors […]. Thus, for students, the value they obtain is either 'right' or 'wrong', which reduces the process to a decision to keep it or throw it away.» (Munier et al., 2013, S. 2757). Die Schülerinnen und Schüler gehen somit oft davon aus, dass es eine Messung ohne Unsicherheit gibt, die beispielsweise durch ein perfektes Messinstrument oder eine sehr exakte Vorgehensweise erreicht werden kann (Hellwig, 2012). Wenn an das Vorhandensein eines 'wahren' Werts geglaubt wird, kann dies bei der Durchführung von experimentellen Messungen verschiedene Konsequenzen haben. Eine Konsequenz ist beispielsweise, dass die Schülerinnen und Schüler Messungen mit dem Ziel wiederholen, ihr Ergebnis (den 'wahren' Wert) zu bestätigen (Lubben & Millar, 1996). Eine weitere Konsequenz ist, dass sie oft darum bemüht sind, die Streuung der erhobenen Daten zu verringern, um sich dem 'wahren' Wert anzunähern (Hellwig, 2012). Zudem können die

Schülerinnen und Schüler das Ziel einer experimentellen Messung darin sehen, das richtige Ergebnis zu finden. Entsprechen die gemessenen Daten nicht ihren Erwartungen, suchen sie den Fehler bei sich selbst statt dass sie erkennen, dass jede Messung von einer Unsicherheit begleitet wird (Fairbrother & Hackling, 1997).

Messwiederholungen und die Bildung eines Mittelwerts erhöhen die Reliabilität der Messung

(Konzept B, vgl. Verweise in Unterkapitel 2.3.2; vgl. z. B. Arnold et al., 2014; Fairbrother & Hackling, 1997; Gott et al., 2003; Hellwig, 2012; Wilhelm & Kunz, 2016)

Wenn die Schülerinnen und Schüler verstehen, dass jede Messung von einer Unsicherheit begleitet wird, dann sollten sie auch verstehen, dass man sich einem 'wahren' Wert nur annähern kann, zum Beispiel durch das Wiederholen von Messungen und das Bilden eines Mittelwerts (vgl. z. B. Arnold et al., 2014; Gott et al., 2003).

Vor allem jüngere Lernende verstehen das Konzept des Wiederholens von Messungen zum Annähern an einen 'wahren' Wert nicht. So fassen vor allem jüngere Lernende zu viele Messdaten als verwirrend auf und bevorzugen darum nur eine Messung respektive Messreihen mit wenigen Daten (vgl. z. B. Masnick & Morris, 2002; Varelas, 1997). Entsprechend wiederholen Schülerinnen und Schüler beim eigenständigen Experimentieren Messungen oft nicht. Wenn sie Messungen wiederholen, tun sie dies als eine Art Routine oder als eine Antwort auf etwas Unerwartetes, zum Beispiel bei Schwierigkeiten bei der Durchführung oder bei einem unerwarteten Ergebnis (Kanari & Millar, 2004). Zudem protokollieren viele Schülerinnen und Schüler die Ergebnisse ihrer Messwiederholungen nicht und / oder führen auch keine weiteren Schritte durch, um anhand ihrer Messwerte zu einem Endergebnis zu gelangen (Kanari & Millar, 2004). Im Vergleich dazu konnten Lubben und Millar (1996) feststellen, dass nur die wenigsten Schülerinnen und Schüler die explizite Frage verneinen, ob eine gegebene Messung wiederholt werden müsse. Daraus kann geschlussfolgert werden, dass Lernende – auch wenn ihnen die Relevanz von Messwiederholungen theoretisch bekannt sein sollte – beim eigenständigen Experimentieren nicht zwingend Messungen wiederholen. Dies konnte auch Heinicke (2012) im Rahmen ihrer Studie mit Studierenden feststellen. So antworteten im Fragebogen die meisten Studierenden bei einem bestimmten Item, dass die Messung zu wiederholen sei. Eine Woche später führten die Probanden ein Experiment durch, welches stark dem Item des Fragebogens ähnelte. Bei der Beobachtung der Probanden beim Experimentieren

stellte sich heraus, dass nur sehr wenige die Messung tatsächlich wiederholten. Heinicke (2012) beschäftigte sich mit dieser Diskrepanz und stellte einen deutlichen Unterschied zwischen verhaltensfernen Argumentationen im Rahmen von Befragungen und verhaltensnahem Handeln beim tatsächlichen Experimentieren fest. Wenn Schülerinnen und Schüler Messungen wiederholen, gibt es eine Vielfalt an Begründungen dafür. Vor allem bei jüngeren Lernenden wird am häufigsten die Ansicht geäussert, dass Messungen zur Bestätigung eines Ergebnisses wiederholt werden (Lubben & Millar, 1996). Bei älteren Schülerinnen und Schülern beziehen sich die meisten Begründungen auf die Streuung der Messergebnisse (Lubben & Millar, 1996). Einige Schülerinnen und Schüler erklären zum Beispiel, dass der Mittelwert einer Messreihe ein genaueres Ergebnis darstellt als ein einzelner Messwert. Was unter ‘genau’ verstanden wird und wie die Anzahl der Messwerte zur Erhöhung der Genauigkeit beiträgt, bleibt jedoch unklar (Lubben & Millar, 1996). Heinicke (2012) führt weitere Gründe für das Durchführen von Messwiederholungen auf, wie zum Beispiel: (1) Messungen werden wiederholt, weil das immer so gemacht wird; (2) Messungen werden wiederholt, wenn das Ergebnis zu stark von den Erwartungen abweicht; (3) Messungen werden wiederholt, um einen Mittelwert zu berechnen oder (4) wie genau die Messung durchgeführt wurde und welches Vertrauen dem Messwert entgegengebracht wird, entscheidet darüber, ob die Messung wiederholt wird. Des Weiteren haben Schülerinnen und Schüler oft Schwierigkeiten, die Daten der Messwiederholungen in Vereinbarung miteinander zu bringen (vgl. z. B. Deardorff, 2001; Hellwig, 2012). So zeigen die Schülerinnen und Schüler oft Schwierigkeiten dabei, anhand einer Messreihe zu einem Endresultat zu gelangen. Einige Schülerinnen und Schüler führen beispielsweise keine weiteren Schritte durch, um anhand ihrer Messreihe zu einem Endresultat zu gelangen oder sie sehen sogar davon ab, die aus den Messwiederholungen resultierenden Werte zu protokollieren (Kanari & Millar, 2004). Insofern mit den Daten der Messwiederholungen weitergearbeitet wird, können verschiedene Strategien seitens der Schülerinnen und Schüler beobachtet werden:

- Die Werte der Messwiederholungen werden genutzt, um vorangehende Messwerte zu ersetzen (Kanari & Millar, 2004).
- Der bestätigte Wert wird als Schlussergebnis gewählt (Lubben & Millar, 1996).
- Der überzeugendste Wert wird als Schlussresultat gewählt (Coelho & Séré, 1998).
- Der erste oder letzte Wert der Messreihe wird als Schlussergebnis genommen (Coelho & Séré, 1998).
- Es wird ein Intervall als Ergebnis angegeben (Coelho & Séré, 1998).

- Es wird ein Mass der zentralen Tendenz (Modus, Median oder Mittelwert) als Schlussergebnis angegeben (vgl. z. B. Coelho & Séré, 1998; Varelas, 1997). Dabei können der Modus und Median als Vorläufer eines statistischen Verständnisses des Mittelwerts angesehen werden (Coelho & Séré, 1998).

Für eine genaue Messung muss ein möglichst genaues Messinstrument verwendet werden

(Konzept C, vgl. Verweise in Unterkapitel 2.3.2; vgl. z. B. Gott et al., 2003; Schulz, 2021; Wilhelm & Kunz, 2016)

Eine wichtige Fähigkeit beim naturwissenschaftlichen Messen ist die Wahl und Verwendung eines geeigneten Messinstruments zum Erfassen der gesuchten Grösse. Hierfür müssen die Schülerinnen und Schüler grundsätzliche Prinzipien von Messinstrumenten verstehen (Gott et al., 2003): Sie müssen beispielsweise wissen, wie sie das Messinstrument bedienen können (Handhabung von Messinstrumenten), wie die gesuchte Grösse auf der Skala abgelesen werden kann und wie die Skala des Messinstruments zu interpretieren ist. Zudem sollten die Schülerinnen und Schüler auch verstehen, dass es kein perfektes Messinstrument gibt, sondern dass auch jedes Messinstrument von einer Messunsicherheit begleitet wird (Gott et al., 2003).

Der Umgang mit Messinstrumenten ist anspruchsvoll und bei der Nutzung können verschiedene Fehler unterlaufen (Haag et al., 2018). Eine Schwierigkeit besteht darin, dass sich die Schülerinnen und Schüler gleichzeitig auf die Handhabung des Messinstruments und auf das Ablesen der Messwerte, zum Beispiel auf einer analogen Skala, konzentrieren müssen. Beim Ablesen der Messwerte auf der Skala können sich weitere Schwierigkeiten ergeben: Der Messwert muss korrekt abgelesen und die Skala korrekt interpretiert werden, sonst ergibt sich ein Ablesefehler, und bei manchen Messinstrumenten müssen weitere Aspekte berücksichtigt werden (z. B. Blickwinkel beim Ablesen eines Messwerts auf einem analogen Thermometer). Zudem zeigt sich, dass die Schülerinnen und Schüler die Genauigkeit von Messinstrumenten oft nicht in Frage stellen und somit davon ausgehen, dass es ein perfektes Messinstrument ohne Messungenauigkeit gibt: «Many students (between 30 % and 60 %) ‘appeared to think that with good enough apparatus and enough care it is possible to make a perfect measurement of a quantity’» (Munier et al., 2013, S. 2758). Deshalb erkennen die Schülerinnen und Schüler oft das Messinstrument nicht als mögliche Ursache für Messunsicherheiten (Munier et al., 2013).

Mengenvergrösserung kann die Messgenauigkeit erhöhen
(Konzept D, vgl. Verweise in Unterkapitel 2.3.2)

Durch das Messen mit einer Menge kann die Messgenauigkeit erhöht werden. Die Strategie der Mengenvergrösserung kann angewendet werden, wenn angenommen wird, dass der gemessene Wert der Menge der Summe der Messwerte der einzelnen Objekte, Ereignisse oder Vorgänge entspricht (Suida & Grabowski, 2012). Das Messen mit einer Menge dient insbesondere dann der Erhöhung der Messgenauigkeit, wenn angenommen wird, dass die Messung mit beispielsweise nur einem Objekt sehr ungenau wird[6].

Das Konzept der Mengenvergrösserung, um ein möglichst genaues Ergebnis zu erhalten, ist insofern anspruchsvoll, da es ein Denken in Proportionalitäten voraussetzt. Schülerinnen und Schüler haben oft Schwierigkeiten beim Denken in Proportionalitäten und ihnen unterlaufen auch beim Rechnen mit Proportionalitäten häufig Fehler, beispielsweise indem Dividend und Divisor vertauscht werden (Hafner, 2012).

2.3.2 Naturwissenschaftliches Messen im Rahmen vorliegender Arbeit: Der Problemtyp «Messen»

Die in Unterkapitel 2.3.1 beschriebenen Konzepte beim naturwissenschaftlichen Messen sind beim Problemtyp «Messen», der im Rahmen vorliegender Arbeit im Fokus steht, zentral. Beim Problemtyp «Messen» geht es darum, eine Grösse mit vorgegebenen Messinstrumenten möglichst genau zu messen. Hierfür stehen den Schülerinnen und Schülern jeweils zwei verschiedene Messinstrumente zur Verfügung, die sich in der Genauigkeit ihrer Skalen unterscheiden. Ein Ziel ist somit, dass die Schülerinnen und Schüler das genauere Messinstrument für das Messen der gesuchten Grösse erkennen und für ihre Messung wählen (vgl. Konzept C, Unterkapitel 2.3.1). Zentral beim Problemtyp «Messen» ist zudem, dass die Schülerinnen und Schüler beim Messen der gesuchten Grösse Strategien anwenden, um ein möglichst genaues Ergebnis zu erhalten. Hierzu gehört beispielsweise Messungen zu wiederholen und einen Mittelwert zu bilden, um sich dem gesuchten Wert anzunähern (vgl. Konzept B, Unterkapitel 2.3.1) oder das Messen mit einer Menge zur Erhöhung der Messgenauigkeit (vgl. Konzept D, Unterkapitel 2.3.1).

[6] Z. B. *Bestimme möglichst genau die Masse eines Kürbissamens.* Das Messen der Masse eines einzelnen Kürbissamens mit einer analogen Waage wird sehr ungenau. Die Strategie der Mengenvergrösserung (mit mehreren Kürbissamen messen) erhöht die Messgenauigkeit und zudem können unterschiedliche Samengrössen berücksichtigt werden.

Bei der vorliegenden Arbeit wurde aufgrund von durchgeführten Pilotstudien im Rahmen des Projekts ExKoNawi (vgl. z. B. Gut, Metzger, et al., 2014; Metzger et al., 2014) fünf hierarchisch angeordnete Qualitätsstandards (QS) für den Problemtyp «Messen» festgelegt (vgl. Tab. 2). Dabei wird davon ausgegangen, dass Schülerinnen und Schüler zum Beispiel einfacher beziehungsweise häufiger QS 1 als QS 2 respektive QS 2 als QS 3 erreichen, wobei QS 5 als der anspruchsvollste Qualitätsstandard angesehen wird.

Tabelle 2: Qualitätsstandards des Problemtyps «Messen» und kodierte Kriterien

	Beschreibung QS	**Kodierte Kriterien**
QS 1	**Einzelmessung:** Eine zur Problemstellung passende Vorgegensweise entwickeln; die Vorgehensweise durchführen, sodass mind. ein Wert im Toleranzbereich resultiert.	- adäquater Versuchsaufbau (1 P) - mind. ein Messwert innerhalb des Toleranzbereichs (1 P) - mind. einmal das für das Messen der gesuchten Grösse genauere Messinstrument verwendet (1 P)
QS 2	**Daten und Ergebnis:** Messdaten aufnehmen und mind. ein Ergebnis im Toleranzbereich mit korrekter Masseinheit angeben.	- Messwerte passen zur durchgeführten Vorgehensweise (1 P) - eindeutiges Ergebnis im Toleranzbereich (1 P) - das Ergebnis hat die korrekte Einheit (1 P)
QS 3	**Messstrategien:** <u>Messwiederholung (3a):</u> Messungen wiederholen, um ein möglichst genaues Ergebnis zu erhalten. Mit den Daten der Messwiederholung weiterarbeiten (z. B. Mittelwertbildung). ODER <u>Mengenvergrösserung (3b):</u> Mit einer (grossen) Menge messen, um ein möglichst genaues Ergebnis zu erhalten. Den Wert der Menge auf die gesuchte Grösse zurückrechnen.	<u>Messwiederholung:</u> - Daten zu Messwiederholungen sind vorhanden und ein Wert wird als Resultat ausgewählt / berechnet (1 P) - Es wird ein Mittelwert aus den Daten der Messwiederholungen berechnet (1 P) - (noch mehr) Messwiederholungen werden als Lösungsvorschlag zur Steigerung der Messgenauigkeit angegeben (1 P) <u>Mengenvergrösserung:</u> - Messwert für eine Menge ist vorhanden (1 P) - Messwert für eine Menge wird auf die gesuchte Grösse zurückgerechnet (1 P) - Messen mit einer (noch grösseren) Menge wird als Lösungsvorschlag zur Steigerung der Messgenauigkeit angegeben (1 P)

QS 4	**Messinstrument:** Das für das Messen der gesuchten Grösse genauere Messinstrument erkennen und für die Messung wählen.	- Das für das Messen der gesuchten Grösse genauere Messinstrument wird erkannt und die Begründung ist korrekt (1 P) - Für die Lösung wird das für das Messen der gesuchten Grösse genauere Messinstrument verwendet (1 P) - Ein (noch) genaueres Messinstrument für das Messen der gesuchten Grösse wird als Lösungvorschlag zur Steigerung der Messgenauigkeit angegeben (1 P)
QS 5	**Quellen für Messunsicherheiten:** Quellen für Messunsicherheiten und Lösungsvorschläge zur Steigerung der Messgenauigkeit nennen.	- Gründe für genaues respektive ungenaues Messen werden genannt (qualitative Codes für das Nennen von möglichen Gründen) - Lösungsvorschläge zur Steigerung der Messgenauigkeit werden genannt (qualitative Codes für das Nennen von möglichen Lösungsvorschlägen)

In Tabelle 2 wird ersichtlich, dass sich QS 1, QS 2, QS 3 und QS 4 aus mehreren dichotom kodierten Kriterien zusammensetzen. Diese Kriterien wurden mit erfüllt (1 Punkt) respektive nicht erfüllt (0 Punkte) beurteilt. Anhand der Gesamtpunktzahl von QS 1 bis QS 4 wurden die experimentellen Kompetenzen quantitativ diagnostiziert. Ein QS wird als erfüllt betrachtet, wenn eine a priori normativ festgelegte Anzahl der Kriterien erreicht werden. Dafür müssen bei QS 1, QS 2 und QS 4 zwei von drei Punkten erreicht werden. Bei QS 3 wird nur eines der beiden Konzepte (‘Messwiederholung’ oder ‘Mengenvergrösserung’) zum Erreichen des QS erwartet, wobei grundsätzlich immer beide Konzepte möglich sind. Somit ist QS 3 erfüllt, sobald zwei von drei Punkten bei einem der beiden Konzepte erreicht werden. QS 5 hingegen wurde qualitativ ausgewertet und liefert Hinweise darüber, inwiefern die Schülerinnen und Schüler Gründe für Messunsicherheiten erkennen und mögliche Lösungsvorschläge ableiten können. Die Daten von QS 5 wurden im Rahmen der Auswertungen vorliegender Arbeit nicht genutzt, können jedoch für zukünftige Forschung dienen (vgl. Kapitel 9). Das den QS und kodierten Kriterien zugrunde liegende Kodiermanual wird in Unterkapitel 7.1.1 beschrieben.

3. Diagnose experimenteller Kompetenzen

Ausgehend von Modellen experimenteller Kompetenzen können Aufgaben und Testinstrumente entwickelt werden, anhand welcher die experimentellen Kompetenzen diagnostiziert werden können (Kampach, 2018). Bei der Diagnose experimenteller Kompetenzen stellt sich die Frage nach der für den Verwendungszweck geeigneten Testart und Erhebungsmethode. Zur Diagnose experimenteller Kompetenzen gibt es verschiedene Testarten. Zu diesen gehören zum Beispiel schriftliche Tests, Tests mit Computersimulationen oder Tests mit Realexperimenten. Zusätzlich gibt es verschiedene Erhebungsmethoden, anhand welcher die experimentellen Kompetenzen erfasst werden können. Bei Tests mit Realexperimenten können beispielsweise die experimentellen Kompetenzen anhand von Schülerprotokollen, Beobachtungen (z. B. Videos) oder der Rekonstruktion der Denkprozesse von Schülerinnen und Schülern (z. B. aus Interviews oder durch die Methode des Lauten Denkens) diagnostiziert werden. Auf typische Testarten (vgl. Unterkapitel 3.1) und Erhebungsmethoden (vgl. Unterkapitel 3.2) wird in der Folge eingegangen. Am Ende des Kapitels werden abschliessend die beschriebenen Testarten und Erhebungsmethoden miteinander verglichen (vgl. Unterkapitel 3.3).

3.1 Testarten

In diesem Unterkapitel wird auf Merkmale typischer Testarten zur Diagnose experimenteller Kompetenzen eingegangen. Zudem werden Vor- und Nachteile der jeweiligen Testarten erläutert und es wird gezeigt, wo diese eingesetzt werden können.

3.1.1 Schriftliche Tests

Bei schriftlichen Tests zur Diagnose experimenteller Kompetenzen – oft auch als Paper and Pencil Tests bezeichnet – werden meistens Teilaspekte des Experimentierprozesses (z. B. Fragestellungen formulieren, Hypothesen formulieren, Experiment planen oder Daten auswerten) vorgegeben, die dann von den Schülerinnen und Schülern ergänzt oder bewertet werden. Dabei experimentieren die Schülerinnen und Schüler nicht selbst, sondern sie müssen sich in eine fiktive experimentelle Situation hineinversetzen. Dadurch fehlt bei schriftlichen Tests eine direkte Rückkoppelung: Die Schülerinnen und Schüler erfahren nicht, ob ihre Versuchsanordnung funktioniert oder nicht (Schreiber et al., 2014). Bei schriftlichen Tests werden typischerweise geschlossene Aufgaben (Multiple Choice, z. B. Hammann et al., 2007; Koenen, 2014)[7], offene Aufgaben (z. B.

[7] Beispiel: Von verschiedenen Hypothesen diejenige wählen, die anhand eines dargestellten Designs überprüft werden kann.

short-answer questions, Baxter & Shavelson, 1994)[8] oder eine Kombination aus beiden Aufgabenformaten (z. B. Erb & Bolte, 2011) eingesetzt.

Ein Vorteil von schriftlichen Tests ist, dass die Teilaufgaben meistens unabhängig voneinander sind und somit kaum Reihenfolgeeffekte entstehen: Obschon vielleicht die Teilaufgabe zur Experimentplanung falsch gelöst wurde, kann dennoch die Aufgabe zur Datenauswertung korrekt bearbeitet werden. Ein weiterer Vorteil ist, dass schriftliche Tests ohne grossen materiellen und personellen Aufwand durchgeführt werden können und sich somit auch für large-scale Assessments eignen. Ein Nachteil von schriftlichen Tests ist jedoch, dass sie handlungsbezogene experimentelle Kompetenzen, wie beispielsweise die effektive Experimentdurchführung inklusive der Handhabung von Messinstrumenten, nicht ausreichend erfassen und abbilden können (vgl. z. B. Baxter & Shavelson, 1994; Hammann et al., 2008; Schreiber, 2012).

3.1.2 Tests mit Computersimulationen

Bei Tests mit Computersimulationen führen die Schülerinnen und Schüler anhand eines virtuellen Versuchsaufbaus Messungen durch. Die Simulation ermöglicht interaktive Handlungen mit dem Experimentiermaterial, sodass eine direkte Rückkoppelung möglich wird: Die Schülerinnen und Schüler erfahren, inwiefern die selbst 'aufgebaute' Versuchsanordnung funktioniert oder nicht. Hiermit können neue Entscheidungen für das weitere Vorgehen getroffen werden (Schreiber, 2012). Zwar ermöglichen Computersimulationen eine interaktive Handlung mit dem Experimentiermaterial, aber der direkte Umgang mit dem Material, wie zum Beispiel das haptische Wahrnehmen der Materialien und das Handhaben von Messinstrumenten, sind mit Computersimulationen nicht möglich. Bei Computersimulationen besteht die Möglichkeit, dass die Schülerinnen und Schüler selbst ihre Daten in einem Protokoll in digitaler Form festhalten (Schülerprotokoll, Schreiber et al., 2014) oder dass die Verhaltensdaten automatisiert erfasst werden (Log-Files). Tests mit Computersimulationen können produkt- oder prozessbezogen ausgewertet werden. Eine produktorientiere Auswertung basiert auf Produkten, wie zum Beispiel Schülerprotokollen. Was während der Testdurchführung genau passiert (z. B. welche Handlungen in welcher zeitlichen Abfolge durchgeführt werden), fliesst in die produktorientierte Auswertung nicht ein. Durch beispielweise Log-Files können Tests mit Computersimulationen auch prozessorientiert ausgewertet werden. Eine prozessorientierte Auswertung ermöglicht das Betrachten von Handlungen während der Testdurchführung, zum

[8] Beispiel: Zu einem gegebenen Phänomen mögliche Fragestellungen formulieren, die untersucht werden könnten.

Beispiel auch im zeitlichen Verlauf. Diese Handlungen können beispielsweise in Handlungsbildern veranschaulicht werden (vgl. Schreiber, 2012). Computersimulationen werden in zunehmendem Masse eingesetzt, um Lernprozesse im Unterricht zu unterstützen. Studien, die Computersimulationen zur Diagnose experimenteller Kompetenzen eingesetzt haben, wurden bis anhin wenige durchgeführt (Schreiber, 2012). Beispielhafte Studien, die Computersimulationen zur Diagnose experimenteller Kompetenzen eingesetzt haben, sind die von Baxter und Shavelson (1994), Dickmann (2016), Schreiber (2012) und Shavelson und anderen (1991).

Ein Vorteil von Tests mit Computersimulationen ist, dass neben den benötigten Computern der materielle und personelle Aufwand gering ist. Somit eignen sich Tests mit Computersimulationen prinzipiell auch für large-scale Assessments (z. B. internationale Vergleichsstudie TIMSS; vgl. Schwippert et al., 2020). Ein Nachteil von vielen Tests mit Computersimulation ist, dass die einzelnen Teilaufgaben aufeinander aufbauend sind: Wenn zum Beispiel der Versuchsaufbau nicht abgeschlossen wurde, entstehen auch keine Messdaten und keine Auswertung. Dadurch bilden sich Reihenfolgeeffekte[9]. Ein weiterer Nachteil von Tests mit Computersimulationen besteht darin, dass einige handlungsbezogene experimentelle Kompetenzen (z. B. die Handhabung von Messinstrumenten) nicht erfasst und abgebildet werden können.

3.1.3 Tests mit Realexperimenten

Tests mit Realexperimenten – oft auch als hands-on Tests bezeichnet – werden in vielen Studien eingesetzt, da sie den Anforderungen des realen Wissenschaftsbetriebs am nächsten kommen (Arndt, 2016). Tests mit Realexperimenten sind dadurch gekennzeichnet, dass die Schülerinnen und Schüler zur Interaktion mit dem realen System angeregt werden. Somit können sie Erfahrungen im Bereich des Aufstellens der Versuchsanordnung, der Handhabung von Messinstrumenten oder der Generierung von Messdaten machen und können zudem die Materialien auch haptisch wahrnehmen. Bei Tests mit Realexperimenten erfahren die Schülerinnen und Schüler direkt, ob ihr Versuchsaufbau funktioniert oder nicht und können unmittelbar Anpassungen vornehmen. Eine direkte Rückkoppelung ist somit gegeben. Bei der Diagnose experimenteller Kompetenzen mit

[9] Es gibt Studien mit Computersimulationen, die versuchen den Einfluss von Reihenfolgeeffekten zu minimieren. Z. B. Dickmann und andere (2014) minimieren Reihenfolgeeffekte, indem jedes Item mit einer Beispiellösung des vorherigen Items startet. So können Schülerinnen und Schüler, denen es nicht gelingt, einen adäquaten Versuchsaufbau zu entwickeln, dennoch mit dem nächsten Item zum Messen von Daten fortsetzen, weil ihnen eine funktionierende Versuchsanordnung in Form einer Beispiellösung präsentiert wird.

Realexperimenten können die Tests produkt- (z. B. Baxter & Shavelson, 1994; Garden, 1999; Gut et al., 2010; Hammann et al., 2008; Stebler et al., 1998) oder prozessbezogen (z. B. Emden & Sumfleth, 2012; Kirchner & Priemer, 2010; Schreiber, 2012) ausgewertet werden. Eine prozessorientierte Auswertung kann mit Hilfe von Verlaufsprotokollen (z. B. Emden & Sumfleth, 2012, vgl. Unerkapitel 3.2.1), Beobachtungen (z. B. Schreiber, 2012) und / oder der Methode des Lauten Denkens (z. B. Kirchner & Priemer, 2010) erfolgen. Der Verlauf während des Experimentierens kann in sogenannten Prozessgrafiken (z. B. Emden & Sumfleth, 2012; Kirchner & Priemer, 2010) oder Handlungsbildern (z. B. Schreiber, 2012) veranschaulicht werden.

Ein Vorteil von Tests mit Realexperimenten ist, dass die Schülerinnen und Schüler selbstständig ein Experiment durchführen können und somit auch das Aufbauen der Versuchsanordnung oder die Handhabung von Messinstrumenten berücksichtigt werden kann. Darum werden Tests mit Realexperimenten als die ideale Testart angesehen (Wenning, 2007) und die entsprechenden Testergebnisse als Benchmark betrachtet (Baxter & Shavelson, 1994). Dennoch weisen Tests mit Realexperimenten auch Nachteile auf. Ein Nachteil ist, dass Tests mit Realexperimenten im Vergleich zu schriftlichen Tests oder Tests mit Computersimulationen einen grossen materiellen und personellen Aufwand aufweisen und daher oft schwierig in large-scale Assessments zu realisieren sind. Ein weiterer Nachteil von Tests mit Realexperimenten besteht in möglichen Reihenfolgeeffekten: Wenn es einem Schüler oder einer Schülerin nicht gelingt, einen Versuch aufzubauen, können auch keine Messdaten generiert und ausgewertet werden[10].

3.2 Erhebungsmethoden

Neben verschiedenen Testarten kommen bei der Diagnose experimenteller Kompetenzen in Tests mit Computersimulationen oder Realexperimenten auch verschiedene Erhebungsmethoden zum Einsatz, zum Beispiel Schülerprotokolle, Beobachtungen (z. B. auch Videos) oder die Rekonstruktion von Denkprozessen von Schülerinnen und Schülern (z. B. aus Interviews oder durch die Methode des Lauten Denkens). Auf diese Erhebungsmethoden wird in der Folge eingegangen.

[10] Es gibt Studien, die versuchen den Einfluss von Reihenfolgeeffekten bei Tests mit Realexperimenten zu minimieren. Z. B. instruierten Solano-Flores und andere (1999) die verschiedenen Teilaspekte des Experimentierprozesses separat: In einem ersten Teil des Tests werden die Schülerinnen und Schüler aufgefordert ein Experiment zu planen; danach führen sie mit Experimentiermaterialien ein Experiment durch; anschliessend werden den Schülerinnen und Schülern Daten vorgegebenen, die sie auswerten und analysieren müssen und in einem letzten Teil des Tests wenden die Schülerinnen und Schüler Informationen an.

3.2.1 Schülerprotokolle

Bei Schülerprotokollen rapportieren die Lernenden ihre Handlungen und Ergebnisse selbstständig, zum Beispiel bei Tests mit Realexperimenten (z. B. Baxter & Shavelson, 1994; Emden & Sumfleth, 2012; Gut-Glanzmann, 2012; Schreiber, 2012). Die Schülerprotokolle können unterschiedlich stark vorstrukturiert sein und offene (z. B. das Planen und Beschreiben einer Vorgehensweise) und geschlossene Antwortformate (z. B. Multiple Choice) beinhalten. Toh und Woolnough (1990) haben bei Schülerprotokollen verschieden stark vorstrukturierte Formate verglichen: Ein minimal vorstrukturiertes Format, ein teilweise vorstrukturiertes Format[11] und ein maximal vorstrukturiertes Format. Dabei stimmte das teilweise vorstrukturierte Format am besten mit der Diagnose überein, die ein Experte oder eine Expertin durch Beobachten von Schülerinnen und Schülern während des Experimentierens vornahm. Des Weiteren unterscheidet sich der Aufbau von Schülerprotokollen je nachdem ob diese produkt- oder prozessbezogen ausgewertet werden sollen. Werden Schülerprotokolle produktbezogen ausgewertet (z. B. Baxter & Shavelson, 1994; Gut-Glanzmann, 2012), dann werden die Lernenden anhand von Aufträgen im Schülerprotokoll aufgefordert beispielsweise ihre Vorgehensweise, Ergebnisse und Schlussfolgerungen zu notieren. Werden die Schülerprotokolle prozessbezogen ausgewertet, dann wird zum Beispiel auch der zeitliche Verlauf während der Durchführung des Experimentierens berücksichtigt. Ein prozessorientiertes Schülerprotokoll – auch als Verlaufsprotokoll bezeichnet – setzten Emden und Sumfleth (2012) ein. Bei diesem Verlaufsprotokoll wurden die Schülerinnen und Schüler während der Durchführung des Experiments aufgefordert, ihre Handlungen simultan zum Bearbeitungsprozess zu protokollieren. Dazu wurde die zur Verfügung stehende Bearbeitungszeit in mehrere Zeitfenster unterteilt, für welche die Lernenden jeweils einen Protokolleintrag formulieren sollten.

Schülerprotokolle bieten, im Vergleich zu beispielsweise Beobachtungen von Schülerinnen und Schülern während des Experimentierens (vgl. Unterkapitel 3.2.2), eine zeitökonomische Alternative, die sich grundsätzlich auch für large-scale Assessments eignet. Bei Schülerprotokollen stellt sich jedoch die Frage, wie sichergestellt werden kann, dass die Schülerinnen und Schüler protokollieren, was sie gemacht haben (Gut-Glanzmann, 2012). Denn anhand von Schülerprotokollen kann nur eine indirekte Aussage über die anwendungsbezogenen experimentellen Kompetenzen getroffen werden (vgl. z. B. Abrahams et al., 2013).

[11] Ein teilweise vorstrukturiertes Format beschreibt eine Serie von offenen Fragen zu möglichst vielen Aspekten der Experimentieraufgabe.

3.2.2 Beobachtungen

Bei der Beobachtung werden die Schülerinnen und Schüler direkt während des Experimentierens beobachtet (z. B. durch geschulte Raterinnen und Rater) oder es werden Videoaufzeichnungen gemacht und diese nachträglich ausgewertet (Häder, 2010). Bei der Beobachtung können nur experimentelle Handlungen (z. B. bei Einzelarbeit) oder auch verbale Äusserungen (z. B. bei Partner- bzw. Gruppenarbeit) berücksichtigt werden. Zur standardisierten Erhebung von Beobachtungsdaten können Beobachtungsbögen oder Kodiermanuale eingesetzt werden. Diese enthalten verschiedene Kriterien oder Kategorien, anhand welcher die experimentellen Kompetenzen erfasst werden. Bei der Diagnose experimenteller Kompetenzen von Schülerinnen und Schülern oder Studierenden durch Beobachtungen kann die Auswertung produkt- (z. B. Baxter & Shavelson, 1994) oder prozessbezogen (z. B. Emden, 2011; Neumann, 2004; Walpuski, 2006) erfolgen. Baxter und Shavelson (1994) haben beispielsweise Lernende während des Experimentierens in Einzelarbeit von trainierten Raterinnen und Ratern beobachten lassen und haben darauf die Daten produktbezogen ausgewertet (z. B. wurden die Experimente korrekt durchgeführt und kamen die Lernenden zu einem richtigen Ergebnis). Emden (2011) und Walpuski (2006) hingegen haben die experimentellen Kompetenzen anhand von Videoaufzeichnungen prozessbezogen ausgewertet, indem sie anhand der Videoaufnahmen Prozessgrafiken erstellt haben, die den experimentellen Verlauf illustrieren.

Um die experimentellen Kompetenzen beim Durchführen von Realexperimenten zu erfassen, wird die Beobachtung oft als Benchmark betrachtet (vgl. z. B. Baxter & Shavelson, 1994; Gott & Duggan, 2002; Gut-Glanzmann, 2012). Ein Nachteil dieser Methode ist jedoch, dass sie sehr ressourcenintensiv ist und sich darum beispielsweise nicht für large-scale Assessments eignet. Zudem lässt sich das Beobachten von Schülerinnen und Schülern während des Experimentierens auf Klassenebene aufgrund der Gruppengrösse kaum realisieren (Emden & Sumfleth, 2012).

3.2.3 Rekonstruktion der Denkprozesse von Schülerinnen und Schülern

Die Denkprozesse von Schülerinnen und Schülern während des Experimentierens können durch die Methode des Lauten Denkens – oft auch als Think Aloud bezeichnet – rekonstruiert werden. Diese Methode ermöglicht Zugänge zu handlungsbezogenen kognitiven Prozessen (Funke & Spering, 2006): Es werden nicht nur die Ergebnisse der Denkprozesse erfassbar, sondern man erfährt auch mehr darüber, wie die Problemlösung zustande gekommen ist. Dadurch eignet sich das Laute Denken auch für eine prozessbezogene Auswertung. Es gibt verschiedene

Formen des Lauten Denkens, die sich bezüglich des Zeitpunkts, der Strukturierung und der Sozialform unterscheiden. Beim gleichzeitigen Lauten Denken werden die Gedanken beim Lösen der experimentellen Problemstellung ausgesprochen (vgl. Konrad, 2020). Beim nachträglichen Lauten Denken wird über eine Entscheidung oder Erfahrung berichtet, die in der Vergangenheit gemacht oder getroffen wurde. Das nachträgliche Laute Denken wird oft dann eingesetzt, wenn die zu bearbeitende Aufgabe komplex ist, da somit die ganze Konzentration für die Aufgabenbearbeitung bleibt. Eine Möglichkeit des nachträglichen Lauten Denken sind Interviews (Völzke, 2012). Beim nachträglichen Lauten Denken werden oft Erinnerungshilfen angeboten, zum Beispiel Videoaufnahmen von dem Experimentieren. Eine solche Vorgehensweise wird als Stimulated Recall bezeichnet (Konrad, 2020). Des Weiteren gibt es offene und strukturierte Formen des Lauten Denkens. Bei offenen Formen werden keine Vorgaben zur Verbalisierung der Gedanken gemacht, während bei strukturierten Formen das Laute Denken durch Prompts (z. B. konkrete Fragen) angeregt wird (Konrad, 2020). Zudem können sich die Formen des Lauten Denkens bezüglich der Sozialform unterscheiden: Das Laute Denken kann sich auf einen Dialog mit Lernpartnerinnen respektive Lernpartnern beziehen oder allein auf den Gedanken eines Individuums fokussieren. Baxter und andere (1995) nutzen beispielsweise die Methode des gleichzeitigen Lauten Denkens mit Prompts (hier Interviewfragen) um mehr über die kognitive Aktivität von Schülerinnen und Schülern während des Experimentierens zu erfahren. Hierbei haben die Lernenden die Problemstellung in Einzelarbeit bearbeitet. Völzke (2012) wiederum nutzt das gleichzeitige und nachträgliche Laute Denken, um mehr über das wissenschaftliche Denken im Kontext des Experimentierens zu erfahren. Dazu lösten die Schülerinnen und Schüler schriftliche Aufträge in Partnerarbeit und äusserten dabei ihre Gedanken. Anschliessend fanden Interviews statt, wobei die Interviewfragen als Prompts und Videoaufnahmen der Partnerarbeit als Erinnerungshilfen dienten.

Da bei der Methode des Lauten Denkens kognitive Prozesse erfassbar werden, eignet sich diese Methode auch für Validierungszwecke (z. B. Vorholzer et al.; 2016). Ein Nachteil dieser Methode ist jedoch, dass sie sehr zeitintensiv ist und sich somit nicht zur Diagnose experimenteller Kompetenzen in large-scale Assessments eignet. Je nach Form des Lauten Denkens ergeben sich zudem weitere Nachteile. Wird zum Beispiel das gleichzeitige Laute Denken bei Problemlösungen in Einzelarbeit eingesetzt, kann das laute Verbalisieren von Gedanken unnatürlich wirken (Völzke, 2012). Beim nachträglichen Lauten Denken (z. B. in Form von Interviews) kann zudem das Problem der sozialen Erwünschtheit auftreten. Dies bedeutet, dass sich die Antwort nicht am subjektiv geglaubten

‘wahren’ Wert orientiert, sondern an den wahrgenommenen Erwartungen (Stocké, 2019). Ein weiterer Nachteil beim nachträglichen Lauten Denken ist, dass Gedanken verfälscht wiedergegeben werden können, weil zum Beispiel Erinnerungslücken auftreten oder nachfolgende Gedanken aufgrund ihrer Folgerichtigkeit bevorzugt berichtet werden (Ericsson & Simon, 1980). Diese Gefahr besteht beim gleichzeitigen Lauten Denken nicht, da die Gedanken vorab nicht reflektiert, strukturiert oder interpretiert und somit verändert werden können (Konrad, 2020).

Die Ausführungen zu den Testarten und Erhebungsmethoden haben gezeigt, dass es verschiedene Möglichkeiten gibt, um die experimentellen Kompetenzen von Schülerinnen und Schülern zu diagnostizieren. Welche Testart beziehungsweise Erhebungsmethode gewählt wird, hängt einerseits von den zu beantwortenden Forschungsfragen, andererseits von Fragen der Testökonomie und der gewählten Stichprobengrösse ab. Welche Testart und welche Erhebungsmethoden im Rahmen vorliegender Arbeit genutzt wurden, wird in Unterkapitel 6.3 berichtet. In den folgenden Ausführungen werden Testarten und Erhebungsmethoden insofern untereinander verglichen, als dass anhand von beispielhaften Studien aufgezeigt wird, inwiefern die Ergebnisse der Kompetenzerfassung miteinander korrelieren und somit zu ähnlichen Ergebnissen führen.

3.3 Vergleich verschiedener Testarten und Erhebungsmethoden

Es gibt einige Studien, die Tests mit Realexperimenten zur Diagnose experimenteller Kompetenzen mit anderen Testarten vergleichen (z. B. Emden, 2011; Hammann et al., 2008; Schreiber, 2012) und somit untersuchen, inwiefern sich Tests mit Realexperimenten, welche durch einen grossen materiellen Aufwand gekennzeichnet sind, durch andere Testarten substituieren lassen. Beispielsweise vergleichen Emden (2011), Hammann und andere (2008), Schreiber (2012), Shavelson und andere (1999) sowie Webb und andere (2000) die Ergebnisse von Tests mit Realexperimenten mit denjenigen von schriftlichen Tests. Die Ergebnisse sind divers, deuten aber auf geringe bis mittlere Korrelationen zwischen den Ergebnissen der Kompetenzerfassung hin (z. B. Emden, 2011; Hammann et al., 2008; Schreiber, 2012; Shavelson et al., 1999)[12]. Darum wird oft argumentiert, dass schriftliche Tests andere Aspekte zu messen scheinen als Tests mit

[12] Emden (2011): Korrelationen zwischen $r = .2$ und $r = .5$ für verschiedene Stichproben; Hammann und andere (2008): Korrelationen zwischen $r = -.3$ und $r = .7$ für verschiedene Testbereiche; Schreiber (2012): Korrelationen zwischen $r = .4$ und $r = .5$ für verschiedene Testbereiche; Shavelson und andere (1999): Korrelationen zwischen $r = .3$ und $r = .5$ für verschiedene Aufgaben.

Realexperimenten (vgl. Gott & Dugan, 2002) und somit diese Testarten nicht austauschbar sind. Schreiber (2012) führt die geringen Korrelationen unter anderem auf die Merkmale der Testarten zurück. So wird beispielsweise bei schriftlichen Tests keine Rückkoppelung ermöglicht, was die Versuchsplanung im schriftlichen Test anspruchsvoller macht. Andererseits scheinen Aufgaben im Bereich des Auswertens bei Tests mit Realexperimenten anspruchsvoller zu sein – möglicherweise weil bei Tests mit Realexperimenten die Gefahr besteht, fehlerhafte Messwerte zu generieren, die dann keine richtige Lösung oder Begründung ermöglichen, was bei vorgegebenen Messwerten in schriftlichen Tests ausgeschlossen wird (Schreiber, 2012). Weitere Gründe für die geringen Korrelationen zwischen den Ergebnissen von schriftlichen Tests und denjenigen von Tests mit Realexperimenten könnten dem Einfluss von inhaltlichen Themen und aufgabenspezifischem Fachwissen auf die Schülerleistungen zuzuschreiben sein (vgl. Gott & Duggan, 2002; Shavelson et al., 1993). Andererseits fanden Webb und andere (2000) hohe Korrelationen (r = .8) beim Vergleich der Schülerleistungen aus schriftlichen Tests und Tests mit Realexperimenten, sodass gefolgert werden könnte, dass schriftliche Tests und Tests mit Realexperimenten ähnliche Facetten zu messen scheinen und somit austauschbar sind. Jedoch scheinen die hohen Korrelationen auf ähnliche Aufgabenstellungen und einen Wiederholungsbeziehungsweise Trainingseffekt zurückzuführen zu sein. Zudem gibt es einige Studien, welche die Ergebnisse der Kompetenzerfassung von Tests mit Realexperimenten mit denjenigen von Tests mit Computersimulationen vergleichen (z. B. Baxter & Shavelson, 1994; Dickmann, 2016; Rosenquist et al., 2000; Schreiber, 2012; Shavelson et al., 1999). Dabei werden oft geringe bis mittlere Korrelationen zwischen den Schülerleistungen der zwei Testarten beobachtet (z. B. Baxter & Shavelson, 1994; Rosenquist et al., 2000; Schreiber, 2012; Shavelson et al., 1991, 1999)[13]. Die Ergebnisse werden unterschiedlich gedeutet: Shavelson und andere (1991) führen die geringe Korrelation auf Probleme mit der Simulation an sich zurück; Shavelson und andere (1999) folgern, dass Tests mit Realexperimenten prinzipiell durch Tests mit Computersimulationen ersetzt werden können, aber aufgrund der Unbeständigkeit der Schülerleistungen eine Methodentriangulation angebracht sei (Beobachtungen und Schülerprotokolle von Tests mit Realexperimenten und Tests mit Computersimulationen); Baxter

[13] Baxter und Shavelson (1994): Korrelationen zwischen r = .4 und r = .7 für verschiedene Aufgaben; Rosenquist und andere (2000): Korrelationen zwischen r = .4 und r = .6 für verschiedene Messzeitpunkte; Schreiber (2012): Korrelationen zwischen r = .3 und r = .6 für verschiedene Testbereiche; Shavelson und andere (1991): Korrelationen zwischen r = .3 und r = .4 für verschiedene Aufgaben; Shavelson und andere (1999): Korrelationen zwischen r = .4 und r = .6 für verschiedene Aufgaben.

und Shavelson (1994) sowie Schreiber (2012) folgern, dass aufgrund der geringen Korrelationen Tests mit Computersimulationen und Tests mit Realexperimenten nicht das Gleiche zu messen scheinen. Schreiber (2012) argumentiert jedoch, dass Tests mit Computersimulationen zumindest das Potential haben, Tests mit Realexperimenten zu ersetzen, denn qualitative Analysen deuten darauf hin, dass Aufgaben von Tests mit Computersimulationen und Realexperimenten sehr ähnlich bearbeitet werden. Dickmann (2016) hingegen konnte zwischen den Schülerleistungen in Tests mit Realexperimenten und Tests mit Computersimulationen bei einigen Testbereichen mittlere bis hohe Korrelationen beobachten (Korrelationen zwischen $r = .6$ und $r = .9$)[14]. Dies spricht in diesen Bereichen dafür, dass Tests mit Realexperimenten durch Tests mit Computersimulationen ersetzt werden können. Dickmann (2016) verglich jedoch die Ergebnisse der Kompetenzerfassung durch Tests mit Computersimulationen und Tests mit Realexperimenten bei Studierenden, ob diese Ergebnisse auch für eine Substituierbarkeit der Testarten bei Schülerinnen und Schülern sprechen, bleibt unklar.

Einige Studien vergleichen bei Tests mit Realexperimenten verschiedene Erhebungsmethoden (z. B. Baxter et al., 1992, 1995; Emden & Sumfleth, 2012; Shavelson et al., 1991, 1993) und untersuchen somit, inwiefern sich bei Tests mit Realexperimenten die Erhebungsmethoden des Beobachtens oder der Rekonstruktion von Denkprozessen von Lernenden (z. B. aus Interviews oder durch das Laute Denken), welches beides sehr zeitintensive Methoden sind, durch Schülerprotokolle ersetzen lassen. Beim Vergleich der Ergebnisse der Kompetenzerfassung durch Beobachtungen von Lernenden und Schülerprotokolle berichten manchen Studien hohe Korrelationen (Baxter et al., 1992; Shavelson et al., 1991, 1993)[15]: Somit scheinen Schülerprotokolle und Beobachtungen etwas Ähnliches zu Messen und Schülerprotokolle einen akzeptablen Ersatz für die Beobachtung zu bieten (Baxter & Shavelson, 1994). Emden und Sumfleth (2012) haben zudem festgestellt, dass es bezüglich des Zusammenhangs zwischen den Ergebnissen der Kompetenzerfassung durch eine prozessorientierte Protokollmethode (vgl. Unterkapitel 3.2.1) und Beobachtungen während des Experimentierens auf die untersuchte Stichprobe ankommt: Während bei der Gymnasialstichprobe hohe Korrelationen (zwischen $r = .8$ und $r = .9$ für verschiedene Experimente) zwischen den erfassten Leistungen festgestellt werden konnten, war dies bei der Gesamtschulstichprobe vorerst nicht der Fall (Korrelationen zwischen $r = .3$ und

[14] Bei einigen Testbereichen gab es auch nicht signifikante Korrelationen. Dickmann (2016) erklärt, dass diese aufgrund der geringen Varianz in den Leistungen der Studierenden entstanden seien und somit nicht gegen die Substituierbarkeit sprechen.

[15] Baxter und andere (1992) und Shavelson und andere (1991): Korrelation von $r = .8$; Shavelson und andere (1993): Korrelationen zwischen $r = .7$ und $r = .8$ für verschiedene Aufgaben.

$r = .4$ für verschiedene Experimente). Die geringen Korrelationen bei der Gesamtschulstichprobe führten Emden und Sumfleth (2012) auf eine kognitive Überlastung der Schülerinnen und Schüler mit der Protokollmethode zurück. Nach einer ausreichenden Methodengewöhnung (Einführung in die Protokollmethode) konnten daraufhin auch bei der Gesamtschulstichprobe hohe Korrelationen (zwischen $r = .8$ und $r = .9$ für verschiedene Experimente) festgestellt werden. Auf ein ähnliches Problem weisen Gott und Dugan (2002) hin. Die Beobachtung von Schülerinnen und Schülern während des Experimentierens lässt sich nicht bei allen Lernenden durch Schülerprotokolle substituieren: Während einigen Lernenden die Planung, Durchführung und Auswertung des Experiments gelingt, scheitern sie beim Schreiben des Protokolls. Umgekehrt gibt es Lernende, die zwar die 'Regeln' zum Schreiben eines gelungenen Protokolls kennen, dieses baut jedoch auf mangelhaften Daten auf (Gott & Dugan, 2002). Studien, welche die Rekonstruktion von Denkprozessen von Lernenden (z. B. aus Interviews oder dem Lauten Denken) mit Schülerprotokollen vergleichen, gibt es wenige. Eine Studie, die ansatzweise Hinweise hierzu liefert, ist die Studie von Baxter und anderen (1995). Bei der Studie von Baxter und anderen (1995) führten die Schülerinnen und Schüler einen Test mit Realexperiment durch, füllten dabei ein Protokoll aus und wurden zusätzlich während der Testdurchführung interviewt. Die erfassten Leistungen anhand der Schülerprotokolle wurden darauf mit den Erkenntnissen aus den Interviews verglichen: Es wurde untersucht, ob Schülerinnen und Schüler mit einer höheren Leistung anhand der Schülerprotokolle auch qualitativ hochwertigere Aussagen in den Interviews zeigten. Dabei konnte festgestellt werden, dass Schülerinnen und Schüler mit einer höheren Leistung anhand der Schülerprotokolle zum Beispiel die Versuchsdurchführung systematischer planten und qualitativ hochwertigere Erklärungen im Interview zeigten. Ein Korrelationsmass zum Veranschaulichen des Zusammenhangs wurde im Rahmen der Studie jedoch nicht berechnet.

Zusammenfassend kann festgestellt werden, dass Tests mit Realexperimenten in Kombination mit der Erhebungsmethode der Beobachtung von Schülerinnen und Schülern während des Experimentierens als Benchmark betrachtet werden können (vgl. z. B. Baxter & Shavelson, 1994; Wenning, 2007). Mit Blick auf die Testarten ist festzuhalten, dass schriftliche Tests andere Aspekte zu messen scheinen als Tests mit Realexperimenten (vgl. z. B. Emden, 2011; Hammann et al., 2008; Schreiber, 2012; Shavelson et al., 1999), während Tests mit Computersimulationen zumindest das Potential haben, Tests mit Realexperimenten zu substituieren (vgl. z. B. Dickmann, 2016; Schreiber, 2012). Mit Blick auf verschiedene Erhebungsmethoden bei Tests mit Realexperimenten zeigt sich, dass

die Beobachtung von Schülerinnen und Schülern während des Experimentierens sich *prinzipiell* durch Schülerprotokolle ersetzen lässt (vgl. z. B. Baxter et al., 1992; Emden & Sumfleth, 2012; Shavelson et al., 1991, 1993). Es konnte jedoch auch gezeigt werden, dass diese Substituierbarkeit nicht zwingend gegeben sein muss und womöglich auch von der untersuchten Stichprobe respektive dem Leistungsniveau der Schülerinnen und Schüler abhängt (vgl. z. B. Emden & Sumfleth, 2012; Gott und Dugan, 2002). Somit stellt sich im Rahmen vorliegender Arbeit die Frage, inwiefern bei der untersuchten Stichprobe (8. Klasse der Sekundarstufe I, vgl. Unterkapitel 6.2) die experimentellen Kompetenzen durch Schülerprotokolle genau erfasst werden können und ob zusätzliche Beobachtungen respektive Videoaufnahmen während des Experimentierens die Genauigkeit der Diagnostik erhöhen. Die Ausführungen in Kapitel 3 haben zudem gezeigt, dass es gegenwärtig noch unklar scheint, inwiefern Schülerprotokolle auch einen akzeptablen Ersatz für die Rekonstruktion von Denkprozessen von Lernenden (z. B. aus Interviews oder durch die Methode des Lauten Denkens) bieten. Somit zeichnet sich hier ein zentrales Desiderat ab: Im Rahmen vorliegender Arbeit wird untersucht, inwiefern zusätzliche Interviews die Genauigkeit der Kompetenzdiagnose erhöhen und in welcher Hinsicht Schülerprotokolle und Schülerprotokolle zusammen mit Interviews zu vergleichbaren Ergebnissen der Kompetenzdiagnose führen.

4. Die Validität von Testverfahren

Gemäss den Teststandards der Fachverbände AERA, APA und NCME (2014)[16] besteht die Grundidee von Validität darin, theorie- und evidenzbasiert zu beurteilen, inwiefern eine Testwertinterpretation für den jeweiligen Verwendungszweck zulässig ist. Die Frage nach der Validität kann dabei nicht allgemeingültig, sondern nur vor dem Hintergrund einer Interpretation in einem sozialen Kontext und für den jeweiligen Verwendungszweck beantwortet werden, das heisst, Validität ist nicht als Merkmal eines Tests zu verstehen (Kane, 2006). Gehen wir beispielsweise von einem schriftlichen Test zur Mechanik aus, welcher zur Standortbestimmung in verschiedenen Klassen eingesetzt wurde und bei welchem die Schülerinnen und Schüler durchweg geringe Leistungen zeigten. Wurde der Test bei Schülerinnen und Schülern mit einem höheren Schulniveau eingesetzt, welche die Aufgaben lesen und verstehen können, ist die Testwertinterpretation vermutlich ‘eher’ valide, da die Testwerte kognitiv valide Schlüsse bezüglich der Kompetenzen der Lernenden im Bereich der Mechanik zulassen. Wurde der gleiche Test hingegen bei Schülerinnen und Schülern mit niedrigem Schulniveau eingesetzt, welche die Aufgaben vermutlich nicht verstehen und / oder lesen können, dann ist die Testwertinterpretation wahrscheinlich ‘eher’ nicht valide, da die Testwerte nicht die Kompetenzen der Lernenden im Bereich der Mechanik, sondern eher sprachliche Probleme widerspiegeln. Die Frage nach der Validität einer Testwertinterpretation gleicht dabei einem Argumentationsprozess, bei welchem nach Hinweisen für beziehungsweise gegen Validität gesucht wird. Somit kann die Frage nach der Validität einer Testwertinterpretation nicht mit *Ja* oder *Nein* beantwortet werden, sondern nimmt unterschiedliche Ausprägungen an. Des Weiteren können bei der Suche nach Hinweisen für Validität unterschiedliche Validitätsaspekte berücksichtigt werden (vgl. Messick, 1995). Diese Validitätsaspekte stellen dabei keine separaten ‘Arten’ von Validität dar, sondern sind als mögliche Elemente einer umfassenden Validitätsbeurteilung zu verstehen (AREA et al., 2014). Welche Validitätsaspekte beziehungsweise in welchem Umfang diese bei der Validitätsbeurteilung berücksichtigt werden, ist vom jeweiligen Verwendungszweck abhängig (Reynolds et al., 2010). Werden beispielsweise Testwertinterpretationen für Laufbahnentscheidungen verwendet, dann bedarf es mehr Hinweisen, die für die Validität der Testwertinterpretation sprechen, als wenn die gleichen Testwerte zur Standortbestimmung eingesetzt werden. Auf die verschiedenen Validitätsaspekte von Messick (1995) wird in der Folge eingegangen.

[16] AERA: American Educational Research Association; APA: American Psychological Association; NCME: National Council on Measurement in Education.

4.1 Validitätsaspekte nach Messick

Messick (1995) geht von einem umfassenden Validitätskonzept aus. Dabei stellt die Konstruktvalidität den Bezugsrahmen für sämtliche Validitätsbetrachtungen dar. Er definiert sechs verschiedene Aspekte von Konstruktvalidität, die bei einer Validitätsbeurteilung berücksichtigt werden können. Diese Validitätsaspekte sind in Tabelle 3 aufgeführt.

Tabelle 3: Validitätsaspekte nach Messick (1995), Übersetzung nach Leuders (2014)

Validitätsaspekte	**Beschreibung**	**Mögliche Hinweise**
Inhaltliche Validität	Relevanz und Repräsentativität der Testinhalte	Hinweise können zum Beispiel durch Literaturrecherchen, das Analysieren von Lehrplänen und Schulbüchern sowie durch Expertenbefragungen gewonnen werden (Messick, 1995).
Kognitive Validität	Passung der kognitiven Prozesse bei der Kompetenzerfassung zu den intendierten Prozessen	Hinweise können zum Beispiel über das Urteil von Expertinnen und Experten oder durch das Untersuchen des Bearbeitungsprozesses gewonnen werden. Der Bearbeitungsprozess kann beispielsweise mit Hilfe der Methode des Lauten Denkens untersucht werden (Leuders, 2014).
Strukturelle Validität	Passung von theoretischem Modell und psychometrischem Messmodell	Hinweise können zum Beispiel über den Vergleich empirischer Daten mit dem theoretisch angenommen Modell gewonnen werden. Wird beispielsweise angenommen, dass sich experimentelle Kompetenzen anhand von drei Teilkompetenzen modellieren lassen, dann sollte diese Struktur in den empirischen Daten ersichtlich werden.
Verallgemeinbarkeit	Angemessenheit der Verallgemeinerung der Testwertinterpretation über die Zielgruppe, die Messzeitpunkte und die Aufgaben hinaus	Hinweise können zum Beispiel gefunden werden, indem der Anteil konstruktirrelevanter Varianz, beispielsweise durch den Einfluss verschiedener Rater oder Messzeitpunkte, bestimmt wird (Dickmann, 2016).

Externe Validität	Angemessenheit der Zusammenhänge mit vorhandenen Theorien oder Konstrukten	Hinweise können zum Beispiel gefunden werden, indem Korrelationen zwischen Testwerten von Testverfahren berechnet werden, bei denen man vermutet, dass sie dasselbe Konstrukt (konvergente Validität) beziehungsweise unterschiedliche Konstrukte (diskriminante Validität) messen (Dickmann, 2016).
Konsequentielle Validität	Angemessenheit der Testwertinterpretation und der daraus abgeleiteten Konsequenzen	Hinweise können zum Beispiel gefunden werden, indem angenommene Konsequenzen überprüft werden. Wird beispielsweise angenommen, dass Tests mit Realexperimenten positive Auswirkungen auf den Unterricht haben, dann sollte geprüft werden, ob sich diese positiven Effekte auch tatsächlich zeigen und ob die negativen Konsequenzen minimal sind (Messick, 1995).

Abhängig von dem Verwendungszweck der Testwertinterpretation und den damit verbundenen Aussagen, die man treffen möchte, sind manche der Validitätsaspekte einmal mehr und einmal weniger zentral (Kane, 2001). Zudem ist aus praktischer Sicht die Berücksichtigung aller Aspekte kaum möglich und wenn überhaupt, dann nur für sehr breit angelegte Studien (Wolming & Wikstrom, 2010). Im Rahmen vorliegender Arbeit liegt der Fokus auf einer kognitiven Validierung eines Testverfahrens am Beispiel der Aufgaben des Problemtyps «Messen». Darum wird in Unterkapitel 4.2 auf diesen Validitätsaspekt eingegangen. Welche weiteren Validitätsaspekte bei der Gesamtvalidierungsstudie des Projekts ExKoNawi untersucht wurden, wird in Unterkapitel 6.1 erläutert.

4.2 Kognitive Validität

Bei der kognitiven Validität wird untersucht, inwiefern die mit einer Aufgabe intendierten kognitiven Prozesse bei der Bearbeitung der Aufgabe tatsächlich aktiviert und von den Lernenden zum Lösen der Aufgabe verwendet werden (Messick, 1995). Hinweise für kognitive Validität können dabei aus unterschiedlichen Bereichen stammen:

(I) Die intendierten Prozesse werden durch die Aufgabenstellung aktiviert und von den Schülerinnen und Schülern zum Lösen der Aufgabe verwendet.
Ob es in diesem Bereich Hinweise auf kognitive Validität gibt, kann beispielsweise überprüft werden, indem die Schülerinnen und Schüler während oder möglichst unmittelbar nach der Testbearbeitung zum Lauten Denken aufgefordert werden (Messick, 1995). Dabei äussern sie ihre Vorgehensweisen, Überlegungen und Gedanken zur Aufgabe. Die von den Schülerinnen und Schülern geäusserten kognitiven Prozesse können dann mit den intendierten Prozessen verglichen werden. Wenn es einen ‘Match’ zwischen den

kognitiven Prozessen seitens der Schülerinnen und Schüler und den intendierten Prozessen gibt, kann dies als Hinweis für kognitive Validität aus dem Bereich (I) gedeutet werden (vgl. Baxter & Glaser, 1998).

(II) Qualitativ hochwertigere Denkprozesse gehen mit einer besseren Lösung der Aufgabe einher.

Als einen weiteren Hinweis für kognitive Validität führen Baxter und Glaser (1998) einen Zusammenhang zwischen der Qualität der gezeigten kognitiven Prozesse seitens der Schülerinnen und Schüler und ihrer Testleistung auf: Schülerinnen und Schüler mit einer höheren Testleistung (also einer besseren Lösung der Aufgabe) sollten auch qualitativ hochwertigere Denkprozesse zeigen. Falls dies der Fall ist, kann dies als Hinweis für kognitive Validität aus dem Bereich (II) gedeutet werden.

(III) Expertinnen und Experten schätzen die intendierten kognitiven Prozesse bei den Aufgaben als naheliegend ein und erkennen keine Aspekte, welche die kognitive Validität beeinträchtigen könnten.

Hinweise für kognitive Validität können auch über das Urteil von Expertinnen und Experten gewonnen werden (vgl. z. B. Leuders, 2014; Miller & Linn, 2000). So können Expertinnen und Experten zum Beispiel befragt werden, inwiefern sie die intendierten Prozesse bei den Aufgaben als naheliegend einschätzen und ob sie in der Aufgabenstellung Merkmale erkennen, welche die kognitive Validität beeinträchtigen könnten (z. B. unbekannte Begriffe). Werden die intendierten kognitiven Prozesse von den Expertinnen und Experten als naheliegend eingeschätzt und keine Merkmale erkannt, welche die kognitive Validität beeinträchtigen könnten, kann dies als Hinweis für kognitive Validität aus dem Bereich (III) gedeutet werden.

Anhand des Beispiels in Abbildung 1 wird im Folgenden illustriert, wie Hinweise für kognitive Validität aus den Bereichen (I) bis (III) gefunden werden können.

Aufgabe:

Die Schachtel liegt auf dem Tisch und bewegt sich nicht, da …

A	… keine Kräfte auf die Schachtel wirken.	Niveau 2
B	… die Schwerkraft zwar nach unten zieht, der Tisch aber im Weg ist.	Niveau 2
C	… die Schwerkraft die Schachtel auf dem Tisch hält.	Niveau 1
D	… der Tisch die Schachtel mit der gleichen Kraft nach oben drückt, wie die Schwerkraft die Schachtel nach unten zieht.	Niveau 3

Abbildung 1: Ordered Multiple Choice-Aufgabe aus der Studie von Alonzo und Steedle (2009); übersetzt von der Autorin vorliegender Arbeit.

Hinweis aus dem Bereich (I): Die intendierten Prozesse werden seitens der Lernenden aktiviert

Nimmt die Mehrheit der Lernenden bei der Begründung für eine gewählte Antwortoption Bezug zu Konzepten im Bereich von Kraft und Bewegung (z. B. «Ich wähle Antwort D, weil die Schwerkraft auf die Schachtel wirkt und es zu jeder Kraft eine Gegenkraft gibt»), ist dies ein Hinweis darauf, dass die intendierten Prozesse aktiviert wurden. Dabei spielt es keine Rolle, ob die Konzepte aus fachlicher Sicht adäquat sind oder nicht, solange ersichtlich wird, dass über ein intendiertes Konzept nachgedacht wurde. Würde die Mehrheit der Lernenden beispielsweise bei der Begründung zur Wahl einer Antwortoption Aussagen machen wie «Ich wähle Antwort A, denn wenn ein Objekt sich nicht bewegt, dann wirken keine Kräfte auf das Objekt», dann sind die gezeigten Konzepte zwar aus fachlicher Sicht falsch, deuten aber dennoch darauf hin, dass über intendierte Konzepte (Konzepte im Bereich der Mechanik) nachgedacht wurde. Wird hingegen bei der Mehrheit der Lernenden ersichtlich, dass sie beim Lösen der Aufgabe auf nicht intendierte Aspekte Bezug nehmen (z. B. «Ich wähle Antwort D, da dieser Satz am längsten ist»), deutet dies darauf hin, dass die Aufgabe keine kognitiv validen Schlüsse bezüglich der Konzepte der Lernenden im Bereich der Mechanik zulässt und somit die Testwertinterpretation nicht valide ist.

Hinweise aus dem Bereich (II): Qualitativ höhere Prozesse gehen mit einer besseren Lösung einher

Wenn ein Schüler oder eine Schülerin ein qualitativ höheres Verständnis des Konzepts zeigt (z. B. versteht, dass ein Objekt sich nicht bewegt, wenn keine Kräfte auf das Objekt wirken oder die resultierende Kraft Null ist), dann sollte auch eine qualitativ bessere Lösung der Aufgabe resultieren, das heisst, dass eine Antwortoption von höherem Niveau gewählt werden sollte (z. B. Antwortoption D). Im Vergleich dazu sollte ein Schüler oder eine Schülerin mit einem qualitativ niedrigeren Verständnisniveau (z. B. versteht '*nur*', dass auf ein Objekt, das sich nicht bewegt, auch keine Kräfte wirken), auch eine Antwortoption von niedrigerem Niveau wählen (z. B. Antwortoption A). Wenn bei der Mehrheit der Lernenden der Zusammenhang zwischen der Qualität der gezeigten Konzepte und der Wahl des Niveaus der Antwortoption gegeben ist, kann dies als Hinweis für kognitive Validität aus dem Bereich (II) gedeutet werden.

Hinweis aus dem Bereich (III): Expertinnen und Experten schätzen die intendierten Konzepte als naheliegend ein und erkennen keine Aspekte, welche die kognitive Validität beeinträchtigen könnten

Um Hinweise für kognitive Validität aus dem Bereich (III) zu finden, kann beispielsweise eine Expertenbefragung durchgeführt werden. Dabei kann die

Intension der Aufgabe erläutert werden (z. B. die Aufgabe hat zum Ziel, Konzepte im Bereich Kräfte und Bewegung zu aktivieren). Wenn die Expertinnen und Experten die intendierten Konzepte bei der Aufgabe als naheliegend einschätzen, kann dies als Hinweis für kognitiv Validität aus dem Bereich (III) gedeutet werden – in dem Sinne, dass eine Aktivierung der intendierten Konzepte seitens der Schülerinnen und Schüler grundsätzlich möglich scheint. Zusätzlich können die Expertinnen und Experten auch gefragt werden, ob sie bei der Aufgabenstellung Aspekte erkennen, welche die kognitive Validität beeinträchtigen könnten (z. B. sprachliche Hürden wie unbekannte Begriffe).

Beim Betrachten von Vorgehensweisen zur kognitiven Validierung von Testverfahren kann festgestellt werden, dass bei den meisten Studien beim Validierungsprozess Hinweise aus dem Bereich (I) berücksichtigt werden und somit untersucht wird, inwiefern die intendierten Konzepte aktiviert und zum Lösen der Aufgaben verwendet werden (z. B. Dickmann, Eickhorst, Theyβen, et al., 2014; Hadenfeldt et al., 2014; Kröger, 2019; Ruiz-Primo et al., 2001; Vorholzer et al., 2016). Im Vergleich dazu werden zusätzliche Hinweise aus den Bereichen (II) und (III) nur bei manchen Studien berücksichtigt (z. B. Hadenfeldt et al., 2014; Kröger, 2019; Ruiz-Primo et al., 2001). Hadenfeldt und andere (2014)[17] sowie Ruiz-Primo und andere (2001)[18] beispielsweise nutzen bei der Validierung eines Testverfahrens Hinweise für kognitive Validität aus den Bereichen (I) und (II). Neben der Berücksichtigung von Hinweisen aus dem Bereich (I) wurde somit beim Validierungsprozess zusätzlich untersucht, inwiefern ein Zusammenhang zwischen der Qualität der gezeigten Konzepte und der Testleistung besteht (Hinweise aus dem Bereich (II)). Kröger (2019) hingegen nutzt beim Validierungsprozess Hinweise für kognitive Validität aus den Bereichen (I) und (III). Somit wird neben der Suche nach Hinweisen für kognitive Validität aus dem Bereich (I) zusätzlich anhand einer Fragebogenstudie untersucht, inwiefern es bei den Aufgaben Hinweise gibt, welche die kognitive Validität beeinträchtigen könnten (Hinweise aus dem Bereich (III)).

[17] Hadenfeldt und andere (2014) validierten ein Testverfahren mit Ordered Multiple Choice-Aufgaben. Hierfür wiesen sie den Schülerinnen und Schülern anhand ihrer Aussagen im Lauten Denken ein Verständnisniveau zu. Danach untersuchten sie, inwiefern das zugewiesene Verständnisniveau dem Niveau der gewählten Antwortoption im Multiple Choice-Test entspricht.
[18] Ruiz-Primo und andere (2001) untersuchten die kognitive Validität von Concept Maps Techniken. Hierfür prüften sie, ob es einen Zusammenhang zwischen der Qualität der gezeigten kognitiven Prozesse und dem Performance Score gibt. Dabei wurde angenommen, dass Schülerinnen und Schüler, die reflektierter vorgehen und mehr adäquate Erklärungen äussern, auch einen höheren Performance Score erzielen im Vergleich zu Lernenden, die eher durch ‘trial and error’ vorgehen und mehr konzeptionelle Fehler äussern.

Zusammenfassend zeigt Kapitel 4, dass bei der Validierung von Testverfahren unterschiedliche Validitätsaspekte berücksichtigt werden können (Messick, 1995). Im Rahmen vorliegender Arbeit liegt der Fokus auf einer kognitiven Validierung eines Testverfahrens mit Realexperimenten. Da bislang keine abgesicherten Theorien zu ablaufenden Überlegungen während Tests mit Realexperimenten vorliegen (Dickmann, 2016), kann a priori nicht verlässlich bestimmt werden, welche Aufgabenmerkmale dazu führen, dass die intendierten Konzepte seitens der Lernenden aktiviert werden. Somit muss a posteriori untersucht werden, inwiefern durch die Aufgabenstellungen die intendierten Konzepte aktiviert und von den Lernenden zum Lösen der Aufgabe verwendet werden. Hiermit zeigt sich ein zentrales Forschungsziel vorliegender Arbeit ab: Am Beispiel der Aufgaben des Problemtyps «Messen» wird untersucht, inwiefern durch die Aufgabenstellungen die intendierten Konzepte aktiviert und von den Lernenden zum Lösen der Aufgaben verwendet werden und somit Hinweise dafür vorliegen, dass die Aufgaben kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen. Zudem hat Kapitel 4 gezeigt, dass bei einer kognitiven Validierung Hinweise aus den Bereichen (I) bis (III) berücksichtigt werden können und anhand der exemplarisch aufgeführten Studien konnte festgestellt werden, dass bei Validierungsprozessen oft hauptsächlich Hinweise aus dem Bereich (I) beigezogen werden, während nur manche Studien auch Hinweise aus den Bereichen (II) und / oder (III) berücksichtigen. Somit besteht ein weiteres Forschungsziel vorliegender Arbeit darin, ein analytisches Verfahren zur kognitiven Validierung eines Testverfahrens mit Realexperimenten unter Berücksichtigung von Hinweisen für kognitive Validität aus den Bereichen (I), (II) und (III) zu entwickeln und auf seine Eignung zu prüfen.

5. Forschungsfragen

In Kapitel 3 wurden Testarten und Erhebungsmethoden beschrieben und exemplarische Studien zum Vergleich dieser aufgeführt. Tests mit Realexperimenten können als Benchmark zur Diagnose experimenteller Kompetenzen betrachtet werden (vgl. z. B. Baxter & Shavelson, 1994; Wenning, 2007), wobei bei diesen die experimentellen Kompetenzen durch verschiedene Erhebungsmethoden erfasst werden können, so zum Beispiel durch Schülerprotokolle, Beobachtungen (z. B. auch Videos) oder die Rekonstruktion von Denkprozessen von Schülerinnen und Schülern (z. B. aus Interviews). Während Beobachtungen und die Rekonstruktion von Denkprozessen von Lernenden sehr zeitintensiv sind und sich auf Klassenebene aufgrund der Gruppengrösse kaum realisieren lassen (vgl. auch Emden & Sumfleth, 2012), bieten Schülerprotokolle diesbezüglich eine ökonomische Alternative an. Es stellt sich jedoch die Frage, inwiefern die experimentellen Kompetenzen in Tests mit Realexperimenten durch Schülerprotokolle genau erfasst werden können (vgl. z. B. Gott & Dugan, 2002; Gut-Glanzmann 2012), da diese nur eine indirekte Aussage über die handlungsbezogenen experimentellen Kompetenzen zulassen (vgl. z. B. Abrahams et al., 2013), und inwiefern *zusätzliche* Erhebungsmethoden (z. B. Videoaufnahmen oder Interviews) die Genauigkeit der Diagnostik erhöhen. Dieser Frage wird in vorliegender Arbeit nachgegangen, wobei die Erhebungsmethoden Schülerprotokolle (abgekürzt *P*, für *P*rotokolle), Videoaufnahmen (abgekürzt *V*, für *V*ideos) und Interviews (abgekürzt *I*, für *I*nterviews) untersucht werden. In vorliegender Studie wird analysiert, inwiefern *zusätzliche* Erhebungsmethoden zum Schülerprotokoll die Genauigkeit der Diagnostik erhöhen, da die Aufgaben und Aufträge in den vorstrukturierten Schülerprotokollen vorzufinden sind und somit auch die Basis für die Videoaufnahmen während des Experimentierens und die Interviews, bei welchen die Lernenden zu den Aufgaben und Aufträgen befragt wurden, bilden. Deshalb werden die Videoaufnahmen und Interviews nicht isoliert, sondern zusammen mit den Schülerprotokollen betrachtet (also PV und PI). Dabei wird angenommen, dass die experimentellen Kompetenzen am genausten durch eine Methodentriangulation (also PVI) diagnostiziert werden können und somit wird PVI als Benchmark betrachtet. Damit stellt sich die Frage, inwiefern die Ergebnisse der Kompetenzdiagnose durch P, PV und PI bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose an den gesetzten Benchmark (PVI) herankommen und somit einen akzeptablen und ökonomischen Ersatz für diesen bieten können. Am Beispiel der Aufgaben des Problemtyps «Messen» führt dies zu Forschungsfrage 1:

FF1 Inwiefern gibt es bei den Aufgaben des Problemtyps «Messen» systematische Abweichungen zwischen P, PV, PI und PVI (gesetzter Benchmark) bezüglich der Genauigkeit des Ergebnisses der Kompetenzerfassung?

Ziel vorliegender Arbeit ist somit zu untersuchen, inwiefern *zusätzliche* Erhebungsmethoden zum Schülerprotokoll die Genauigkeit der Diagnostik bei Tests mit Realexperimenten erhöhen (P vs. PV und P vs. PI) und inwiefern die Ergebnisse der Kompetenzerfassung durch P, PV und PI bezüglich der Genauigkeit des Ergebnisses der Diagnose an den gesetzten Benchmark (PVI) herankommen. Hierfür werden die Ergebnisse der Kompetenzdiagnose jeweils auf Ebene der Stichprobe und auf Ebene einzelner Schülerinnen und Schülern (Individualebene) verglichen. Eine Betrachtung auf beiden Ebenen ist sinnvoll (vgl. Schreiber, 2012), da verschiedene Faktoren (z. B. Motivation der Lernenden, Messzeitpunkt oder Kontext der Aufgabe) die Ergebnisse der Kompetenzdiagnose bei Tests mit Realexperimenten beeinflussen können (vgl. z. B. Gott & Dugan, 2002; Shavelson et al., 1993). So ist es möglich, dass die Erhebungsmethoden auf Ebene einzelner Lernenden nicht zu vergleichbaren Ergebnissen führen, aber dennoch auf Ebene der Stichprobe (z. B. beim Betrachten der Mittelwerte) austauschbar sind. Ein solcher Befund würde dafür sprechen, dass beim Einsatz des Testinstruments für large-scale Assessments die Erhebungsmethoden substituiert werden können, während dies beim Einsatz des Tests für Zwecke der Individualdiagnostik nicht der Fall wäre.

In einigen Studien werden die Ergebnisse der Kompetenzdiagnose durch Schülerprotokolle und Beobachtungen von Lernenden während des Experimentierens verglichen (z. B. Baxter et al., 1992; Emden und Sumfleth, 2012; Shavelson et al., 1991, 1993). Die Ergebnisse der Studien sind unterschiedlich: Während manche Studien (z. B. Baxter et al., 1992; Shavelson et al., 1991, 1993) feststellen, dass die Diagnose experimenteller Kompetenzen durch Schülerprotokolle und Beobachtungen zu sehr ähnlichen Ergebnissen führen und somit austauschbar sind, stellen Emden und Sumfleth (2012) fest, dass dieser Zusammenhang von der untersuchten Stichprobe abhängt. Während beim leistungsstärkeren Schulniveau (Gymnasialstichprobe) die Erhebungsmethoden zu sehr ähnlichen Ergebnissen bezüglich der Kompetenzdiagnose führen und somit substituierbar sind, ist dies beim tieferen Schulniveau (Gesamtschulstichprobe) erst nach einer ausreichenden Einführung in die Protokollmethode der Fall. Im Rahmen vorliegender Studie wurden die experimentellen Kompetenzen von Schülerinnen und Schülern des 8. Schuljahrs untersucht. Die Lernenden der untersuchten Stichprobe waren alle in der Sekundarstufe I und dem höheren Schulniveau (im

Kanton Zürich Leistungsniveau A[19]). Ausgehend von den Ergebnissen von Emden und Sumfleth (2012) wird folglich angenommen, dass in der untersuchten Stichprobe (leistungsstärkeres Schulniveau) Schülerprotokolle auf Individualebene und Ebene der Stichprobe einen akzeptablen Ersatz für das Beobachten von Schülerinnen und Schülern während des Experimentierens (auch Videos) bieten und somit zusätzliches Beobachten nicht zu einem genaueren Ergebnis der Kompetenzdiagnose führt. Daraus ergeben sich die Hypothesen H1 und H2 (vgl. auch Abb. 2):

H1 *Die Diagnose experimenteller Kompetenzen durch P und PV führen auf Ebene der Stichprobe und Individualebene zu einem sehr ähnlichen Ergebnis der Kompetenzdiagnose und sind somit austauschbar.*

H2 *Die Diagnose experimenteller Kompetenzen durch PI und PVI führen auf Ebene der Stichprobe und Individualebene zu einem sehr ähnlichen Ergebnis der Kompetenzdiagnose und sind somit austauschbar. Somit wird angenommen, dass die Diagnose experimenteller Kompetenzen durch PI bezüglich der Genauigkeit des Ergebnisses der Kompetenzerfassung an den gesetzten Benchmark (PVI) herankommt.*

Es gibt kaum Studien, welche die Ergebnisse der Kompetenzdiagnose in Tests mit Realexperimenten anhand der Rekonstruktion von Denkprozessen von Schülerinnen und Schülern (z. B. aus Interviews) mit den Ergebnissen anhand von Schülerprotokollen vergleichen (ansatzweise die Studie von Baxter und anderen (1995), vgl. Kapitel 3). Es kann jedoch davon ausgegangen werden, dass Schülerprotokolle nur einen eingeschränkten Einblick in die Überlegungen und Gedanken von Schülerinnen und Schülern während des Experimentierens gewährleisten, da nicht allen Schülerinnen und Schüler das Führen eines Protokolls gelingt (vgl. z. B. Gott & Dugan, 2002; Gut-Glanzmann, 2012) und zudem Studien darauf hinweisen, dass zwischen schriftlichen Materialien und den Gedanken von Schülerinnen und Schülern nicht stets ein systematischer Zusammenhang besteht (vgl. z. B. Vorholzer et al., 2020). Somit wird angenommen, dass zusätzliche Interviews auf Ebene der Stichprobe und auf Individualebene zu einem genaueren Ergebnis der Kompetenzdiagnose führen. Daraus ergeben sich die Hypothesen H3 bis H5 (vgl. auch Abb. 2):

[19] In der Schweiz werden die Schülerinnen und Schüler ab dem 7. Schuljahr in verschiedene Leistungsniveaus eingeteilt. Dabei stellt das Gymnasium das höchste Leistungsniveau dar. Neben dem Gymnasium gibt es die Sekundarstufe I. Diese wird im Kanton Zürich, in dem die Studie durchgeführt wurde, in die Niveaus A, B und C unterteilt, wobei das Niveau A das höchste Leistungsniveau der Sekundarstufe I ist. In einigen Gemeinden werden die Leistungsniveaus B und C in gemeinsamen Klassen geführt und somit gibt es Klassen der Sekundarstufe I mit Niveau A und Niveau BC.

H3 *Die Diagnose experimenteller Kompetenzen durch PI führt auf Ebene der Stichprobe und auf Individualebene zu einem genaueren Ergebnis der Kompetenzerfassung im Vergleich zu P.*

H4 *Die Diagnose experimenteller Kompetenzen durch PVI führt auf Ebene der Stichprobe und Individualebene zu einem genaueren Ergebnis der Kompetenzerfassung im Vergleich zu P. Somit wird davon ausgegangen, dass P bezüglich der Genauigkeit des Ergebnisses der Kompetenzerfassung keinen akzeptablen Ersatz für den Benchmark (PVI) bietet.*

H5 *Die Diagnose experimenteller Kompetenzen durch PVI führt auf Ebene der Stichprobe und Individualebene zu einem genaueren Ergebnis der Kompetenzerfassung im Vergleich zu PV. Somit wird davon ausgegangen, dass auch PV keinen akzeptablen Ersatz für die Erfassung experimenteller Kompetenzen anhand des Benchmarks (PVI) bietet.*

In Abbildung 2 wird der Vergleich der verschiedenen Erhebungsmethoden illustriert und es werden die Hypothesen H1 bis H5 verortet.

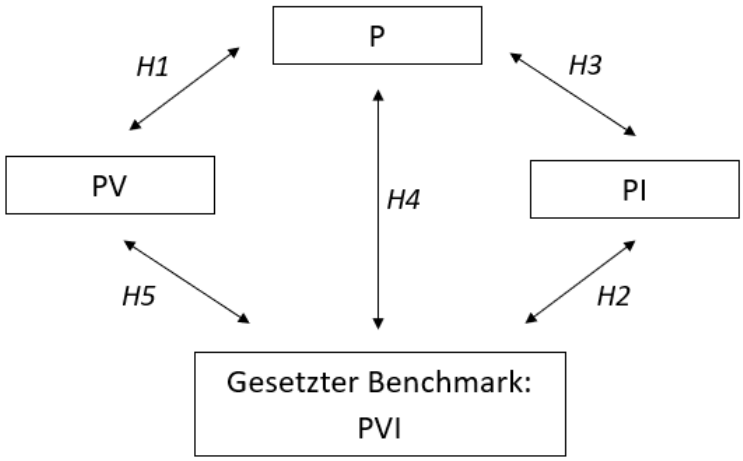


Abbildung 2: Vergleich der Erhebungsmethoden im Rahmen von Forschungsfrage 1 (P = Schülerprotokolle; V = Beobachtungen bzw. Videoaufnahmen; I = Interviews) und Verortung der Hypothesen H1 bis H5.

Beim Erfassen von Kompetenzen stellt sich stets die Frage nach der Validität der Diagnose. Im Rahmen vorliegender Studie wird die kognitive Validität eines Testverfahrens mit Realexperimenten untersucht. In Kapitel 4 wurde erläutert, dass Hinweise für kognitive Validität aus den Bereichen (I) bis (III) stammen können: (I) Die intendierten Konzepte werden aktiviert und zum Lösen der Aufgaben verwendet, (II) qualitativ höhere Denkprozesse gehen mit einer besseren Lösung der Aufgabe einher und (III) Expertinnen und Experten schätzen die intendierten Konzepte als naheliegend ein und erkennen keine Aspekte, welche die

kognitive Validität beeinträchtigen könnten. Am Beispiel der Aufgaben des Problemtyps «Messen» ergibt sich daraus Forschungsfrage 2:

FF2 Inwiefern gibt es bei den Aufgaben des Problemtyps «Messen» Hinweise für kognitive Validität aus den Bereichen (I), (II) und (III)?

Da bereits einige Validierungsstudien im Rahmen des Projekts ExKoNawi stattgefunden haben (vgl. z. B. Gut et al., 2017; Gut, Metzger, et al., 2014; Hild et al., 2017; Metzger et al., 2014) und im Laufe dieser die Aufgaben und Aufträge in den vorstrukturierten Schülerprotokollen stets angepasst und verbessert wurden, wird vermutet, dass Hinweise für kognitive Validität aus den Bereichen (I) bis (III) bei den Aufgaben des Problemtyps «Messen» gefunden werden können. Ziel vorliegender Arbeit ist es somit, entsprechende Hinweise anhand eines analytischen Verfahrens aufzudecken.

Die in Kapitel 4 exemplarisch aufgeführten Studien zur kognitiven Validierung von Testverfahren haben gezeigt, dass meistens im Rahmen von kognitiven Validierungsprozessen Hinweise aus dem Bereich (I) beigezogen werden, während nur manche Studien auch Hinweise aus den Bereichen (II) und / oder (III) berücksichtigen. Somit besteht ein weiteres Ziel vorliegender Arbeit darin, im Zuge von Forschungsfrage 2 zu untersuchen, inwiefern sich die angewandte Vorgehensweise eignet, um ein Testverfahren mit Realexperimenten unter Berücksichtigungen von Hinweisen aus den Bereichen (I), (II) und (III) kognitiv zu validieren.

Im folgenden empirischen Teil der Arbeit (Kapitel 6 bis 8) werden das Untersuchungsdesign sowie die Auswertungen und Ergebnisse im Rahmen des Untersuchens von Forschungsfrage 1 und 2 vorgestellt.

Empirischer Teil

6. Untersuchungsdesign

Die vorliegende Arbeit wurde im Rahmen der Gesamtvalidierungsstudie des Projekts ExKoNawi realisiert und gliedert sich in zwei Teilstudien: Teilstudie I befasst sich mit dem Vergleich der Ergebnisse der Kompetenzdiagnose anhand verschiedener Erhebungsmethoden (FF1), Teilstudie II mit der kognitiven Validierung eines Testverfahrens mit Realexperimenten (FF2) – jeweils am Beispiel des Problemtyps «Messen». Im Folgenden wird zuerst erläutert, wie Teilstudie I und II in die Gesamtvalidierungsstudie des Projekts eingebettet sind (vgl. Unterkapitel 6.1). Die zwei Teilstudien der vorliegenden Arbeit sind aufeinander aufbauend und es werden teilweise die gleichen Daten genutzt: Für die kognitive Validierung des Testverfahrens (Teilstudie II) werden unter anderem die Daten von Teilstudie I (Schülerprotokolle und Interviews) genutzt, wobei diese je nach Forschungsfrage unterschiedlich ausgewertet werden. Daher werden in den folgenden Unterkapiteln (Unterkapitel 6.2 bis 6.4) die allgemeinen Grundlagen beschrieben, die für Teilstudie I und II identisch sind: die Stichprobe, der Ablauf der Datenerhebung sowie die dafür verwendeten Instrumente.

6.1 Einordnung von Teilstudie I und II in die Gesamtvalidierungsstudie

Teilstudie I und II sind in die Gesamtvalidierungsstudie des Projekts ExKoNawi eingebunden. Daraus ergeben sich für diese Arbeit einige Konsequenzen: (1) Die vorliegende Stichprobe ist ein Teil der Gesamtstichprobe (vgl. Abb. 3 und Unterkapitel 6.2); (2) die vorliegenden Daten wurden zu einem erheblichen Teil während der Gesamterhebung erfasst, wodurch sich die Struktur des Ablaufs der Datenerhebung ergab (vgl. Unterkapitel 6.3) und (3) ein Teil der Daten wurde sowohl für die Auswertung im Rahmen der strukturellen und externen Validierung (Dissertationsprojekt Bonetti; vgl. Bonetti, in Vorbereitung) als auch für die Auswertungen im Rahmen vorliegender Arbeit genutzt (z. B. die von den 27 Jugendlichen der Stichprobe ausgefüllten Schülerprotokolle zu den Aufgaben des Problemtyps «Messen» und deren Ergebnisse in den Begleittests).

Abbildung 3: Einordnung von Teilstudie I und II in die Gesamtvalidierungsstudie des Projekts ExKoNawi.

Zentrales Forschungsinteresse der Gesamtvalidierungsstudie ist es, das Testverfahren des Projekts ExKoNawi unter Berücksichtigung verschiedener Validitätsaspekte zu validieren. Das Testverfahren vorliegender Arbeit, sowie des Projekts ExKoNawi, soll dazu dienen, die experimentellen Kompetenzen von Schülerinnen und Schülern der 8. Klasse bei Tests mit Realexperimenten zu diagnostizieren. Ziel ist, dass die Testwerte genutzt werden können, um Aussagen über die Kompetenzen von Schülerinnen und Schülern beim Bearbeiten von Aufgaben mit Realexperimenten zu treffen, im Sinne einer Standortbestimmung. Zur Validierung werden folgende Validitätsaspekte berücksichtigt (vgl. Abb. 3):

- Strukturelle Validierung: Anhand des Testverfahrens sollen die experimentellen Kompetenzen von Jugendlichen diagnostiziert werden, wobei davon ausgegangen wird, dass diese verschiedene Teilkompetenzen umfassen. Die Teilkompetenzen werden als die Fähigkeit aufgefasst, unterschiedliche experimentelle Problemstellungen lösen zu können und entsprechen somit den Problemtypen «Vergleichen», «Untersuchen» und «Messen» (vgl. Unterkapitel 2.2.1). Folglich muss geprüft werden, ob die theoretischen Annahmen zur Kompetenzmodellierung anhand empirischer Daten bestätigt werden können, in dem Sinne, dass sich die Teilkompetenzen (Problemtypen) in der Empirie als differenzierbare Dimensionen erfassen lassen (Dissertationsprojekt Bonetti; vgl. Bonetti, in Vorbereitung).
- Externe Validierung: Das Testverfahren soll die experimentellen Kompetenzen von Jugendlichen erfassen. Somit muss geprüft werden, inwiefern das Testverfahren bei der Kompetenzdiagnose hauptsächlich die experimentellen Kompetenzen von Jugendlichen misst und nicht zu einem erheblichen Teil auch andere Aspekte (z. B. sprachliche Fähigkeiten) mit abgebildet werden (Dissertationsprojekt Bonetti; vgl. Bonetti, in Vorbereitung).
- Kognitive Validierung: Zudem muss geprüft werden, ob zum Lösen der Aufgaben hauptsächlich die intendierten Konzepte verwendet werden. Dadurch kann sichergestellt werden, dass die Ergebnisse der Kompetenzdiagnose kognitiv valide Schlüsse bezüglich der Kompetenzen der Lernenden im Bereich der intendierten Konzepte (z. B. experimentelle Strategien wie das Durchführen von Messwiederholungen) zulassen und nicht andere Strategien (z. B. Raten) zu einem erheblichen Teil das Ergebnis der Kompetenzdiagnose beeinflussen. Dies wird im Rahmen vorliegender Arbeit exemplarisch anhand der sechs Aufgaben des Problemtyps «Messen» untersucht (Teilstudie II, vgl. Kapitel 8).

Falls im Rahmen des Validierungsprozesses Hinweise für die Validität des Testverfahrens gefunden werden können, legitimiert dies den zukünftigen Einsatz des Testinstruments für die Diagnose experimenteller Kompetenzen von

Jugendlichen des 8. Schuljahres und den definierten Verwendungszweck (Standortbestimmung).

Ein weiteres Forschungsinteresse besteht darin, die Ergebnisse der Diagnose experimenteller Kompetenzen bei Tests mit Realexperimenten anhand verschiedener Erhebungsmethoden zu vergleichen (vgl. Abb. 3, Teilstudie I). Schülerprotokolle bieten eine ökonomische Möglichkeit zur Erfassung experimenteller Kompetenzen bei Tests mit Realexperimenten, die sich auch für grosse Stichproben (vgl. Abb. 3, Gesamtstichprobe) und large-scale Assessments eignet. Es stellt sich jedoch die Frage, inwiefern anhand von Schülerprotokollen die experimentellen Kompetenzen in Tests mit Realexperimenten genau erfasst werden können, da diese nur eine indirekte Aussage über die experimentellen Kompetenzen von Lernenden zulassen (vgl. z. B. Abrahams et al., 2013; Gott & Dugan, 2002). Folglich wird im Rahmen von Teilstudie I untersucht, inwiefern die experimentellen Kompetenzen anhand von Schülerprotokollen genau erfasst werden können und somit Schülerprotokolle eine genaue Aussage über die experimentellen Kompetenzen von Jugendlichen zulassen (vgl. Kapitel 7). Falls gezeigt werden kann, dass Schülerprotokolle eine genaue Diagnose ermöglichen, können beim zukünftigen Einsatz dieses Testinstruments die experimentellen Kompetenzen anhand von Schülerprotokollen ökonomisch, aber dennoch genau, erfasst werden.

6.2 Stichprobe

Die Stichprobe von Teilstudie I und II umfasst insgesamt 27 Schülerinnen und Schüler (13 Mädchen und 14 Knaben) des 8. Schuljahres der Sekundarstufe I im Leistungsniveau A (vgl. Abb. 3). Der Fokus auf die Sekundarstufe I, Niveau A wurde gewählt, da einerseits die Stichprobengrösse im Rahmen der Arbeit beschränkt war und somit zusätzliche Varianz aufgrund des Leistungsniveaus vermieden werden sollte und andererseits das Niveau A das mittlere Leistungsniveau der Gesamtstichprobe darstellt (Gesamtstichprobe: Gymnasium und Klassen der Sekundarstufe I mit Niveau A und BC; vgl. Abb. 3 und Fussnote 19). Die Lernenden stammten aus drei verschiedenen Schulen und sechs verschiedenen Klassen (vgl. Tab. 4). Die Wahl der Schulen, Klassen sowie Schülerinnen und Schülern wurde von weiteren Faktoren beeinflusst: (1) die äusseren Rahmenbedingungen mussten für das Durchführen der Videoaufnahmen und Interviews gegeben sein (z. B. genügend Personen für die Datenerhebung) und (2) die Lehrpersonen und Schülerinnen und Schüler mussten sich für die Videostudie und Interviews zur Verfügung stellen (z. B. Einverständniserklärungen der Erziehungsberichtigten). Die Erhebungen fanden in Halbklassen statt, wobei pro

Halbklasse null bis drei Schülerinnen und Schüler untersucht wurden. Aus organisatorischen Gründen (z. B. Rotation der Lernenden zwischen verschiedenen Aufgaben und Anzahl Personen, welche die Videoaufnahmen machen konnten) konnten pro Halbklasse maximal drei Lernende untersucht werden. In manchen Halbklassen wurden keine Schülerinnen und Schüler gefilmt und interviewt, da entweder keine Einverständniserklärung zur Videoaufnahme vorlagen oder die äusseren Rahmenbedingungen nicht gegeben waren (z. B. nicht ausreichend Personen für die Datenerhebung). Jede Halbklasse wurde vier Mal besucht. Bei den vier Messzeitpunkten lösten die Schülerinnen und Schüler unter anderem je eine Aufgabe des Problemtyps «Messen». Somit liegen von den 27 Lernenden der Stichprobe jeweils zu vier Aufgaben des Problemtyps «Messen» Daten vor und folglich gibt es insgesamt 108 Schülerprotokolle, Videoaufnahmen und Interviews zu den Aufgaben des Problemtyps «Messen». Bei der Verteilung der sechs Aufgaben des Problemtyps «Messen» auf die Schülerinnen und Schüler wurde zudem darauf geachtet, dass jede Aufgabe von 18 Lernenden bearbeitet wurde.

Tabelle 4: Beschreibung der Stichprobe

Schule	Zeitraum	Klasse	Halbklasse	untersuchte Anzahl Lernende	Gelöste Aufgaben des Problemtyps «Messen»[20]					
					Ahorn	Bohne	Filzstift	Faden	Münze	Pulver
I	Februar / März 2017	1	a	3	x	x			x	x
			b	3			x	x	x	x
II	März / April 2017	2	a	0						
			b	3		x	x	x	x	
		3	a	3	x		x	x		x
			b	0						
III	Juni 2017	4	a	3	x	x		x		x
			b	0						
		5	a	3	x			x	x	x
			b	3	x	x	x		x	
		6	a	3		x	x		x	x
			b	3	x	x	x	x		
				Σ = 27	**Σ = 18**	**Σ = 18**	**Σ = 18**	**Σ = 18**	**Σ = 18**	**Σ = 18**
					Σ = 108					

[20] Die Aufgaben des Problemtyps «Messen» werden in Unterkapitel 6.4.1 beschrieben.

6.3 Ablauf der Datenerhebung

Im Rahmen der Erhebung fanden pro Halbklasse insgesamt vier Besuche innerhalb eines Zeitraums von zwei bis vier Wochen[21] statt (vgl. Tab. 5). Bei jedem Besuch lösten die Schülerinnen und Schüler drei Aufgaben mit Realexperimenten (Gesamtvalidierungsstudie des Projekts, vgl. Unterkapitel 6.1), davon jeweils eine Aufgabe des Problemtyps «Messen». Während des Lösens der Aufgaben haben die Schülerinnen und Schüler ihre Ergebnisse, Gedanken und Schlussfolgerungen in vorstrukturierten Schülerprotokollen (P) notiert. Im Rahmen vorliegender Studie wurden die Schülerinnen und Schüler zudem während des Lösens der Aufgaben des Problemtyps «Messen» videografiert (V) und zusätzlich nach jedem Besuch – in der Regel direkt anschliessend[22] – interviewt (I). In den Interviews wurden die Schülerinnen und Schüler zu ihren Vorgehensweisen und Ergebnissen befragt und aufgefordert, ihre Gedanken und Überlegungen zu verbalisieren. Als Erinnerungshilfen bei den Interviews standen die von den Schülerinnen und Schülern ausgefüllten Protokolle zur Verfügung. Zudem fanden bei jedem Besuch weitere Begleittests statt. Diese dienten hauptsächlich dazu, im Rahmen der Gesamtvalidierungsstudie die externe Validität des Testverfahrens zu prüfen (vgl. Unterkapitel 6.1). Die Ergebnisse der Begleittests wurden im Rahmen vorliegender Arbeit aber auch qualitativ genutzt, um Unterschiede in den Ergebnissen der Kompetenzdiagnose durch die verschiedenen Erhebungsmethoden ansatzweise qualitativ zu begründen (vgl. Unterkapitel 7.3).

[21] Die Zeitspanne wurde durch äussere Rahmenbedingungen (z. B. Stundenpläne der Schülerinnen und Schüler und unterrichtsfreie Tage) bestimmt.

[22] Bei 84 von 108 geführten Interviews (≙ 78 % der Fälle) fanden die Interviews am gleichen Tag statt. In manchen Fällen war das aus organisatorischen Gründen nicht möglich (z. B. weil der Besuch in den letzten Lektionen des Schultags stattfand und die Lernenden anschliessend unterrichtsfreie Zeit hatten). In diesen Fällen wurden die Interviews in den nächsten ein bis zwei Tagen durchgeführt.

Tabelle 5: Ablauf der Datenerhebung

		Besuch 1	**Besuch 2**	**Besuch 3**	**Besuch 4**
Doppellektion (90 min)	**Realexperimente**	3 Aufgaben, davon 1 Aufgabe des Problemtyps «Messen» *erhobene Daten: P und V*	3 Aufgaben, davon 1 Aufgabe des Problemtyps «Messen» *erhobene Daten: P und V*	3 Aufgaben, davon 1 Aufgabe des Problemtyps «Messen» *erhobene Daten: P und V*	3 Aufgaben, davon 1 Aufgabe des Problemtyps «Messen» *erhobene Daten: P und V*
	Begleittests	Schriftlicher Experimentiertest (Koenen, 2014; Mannel, 2011) Fragebogen Motivation / Situationales Interesse (vgl. Bonetti, in Vorbereitung)	Demographischer Fragebogen (vgl. Bonetti, in Vorbereitung) Fragebogen Motivation / Situationales Interesse (vgl. Bonetti, in Vorbereitung)	KFT (Teilbereiche Q und N; Heller & Perleth, 2000) Fragebogen Motivation / Situationales Interesse (vgl. Bonetti, in Vorbereitung)	KFT (Teilbereich V; Heller & Perleth, 2000); SLS (Mayering & Wimmer, 2014) Fragebogen Motivation / Situationales Interesse (vgl. Bonetti, in Vorbereitung)
Anschliessend		Einzelinterview zur Aufgabe des Problemtyps «Messen» *erhobene Daten: I*	Einzelinterview zur Aufgabe des Problemtyps «Messen» *erhobene Daten: I*	Einzelinterview zur Aufgabe des Problemtyps «Messen» *erhobene Daten: I*	Einzelinterview zur Aufgabe des Problemtyps «Messen» *erhobene Daten: I*

Anmerkungen: Die Abkürzung KFT steht für kognitiver Fähigkeitstest. Dieser besteht aus den Bereichen Q (quantitative (numerische) Fähigkeiten), N (anschauungsgebundenes (figurales) Denken) und V (sprachliches Denken). SLS steht für Salzburger Lese-Screening für die Schulstufen 2–9.

6.4 Instrumente

In der Folge werden die Instrumente beschrieben, die im Rahmen von Teilstudie I und II hauptsächlich genutzt wurden: Die Aufgaben des Problemtyps «Messen», die Schülerprotokolle, die Videoaufnahmen während des Experimentierens und die Interviews zu den Aufgaben.

6.4.1 Aufgaben des Problemtyps «Messen»

Im Rahmen vorliegender Arbeit werden wie im Projekt ExKoNawi Aufgaben mit Realexperimenten zur Diagnose experimenteller Kompetenzen genutzt (vgl. Unterkapitel 3.1.3). Die Schülerinnen und Schüler erhalten eine experimentelle Problemstellung, die sie selbstständig mit vorgegebenen Materialien in 18 Minuten Einzelarbeit lösen. Bei den Aufgaben des Problemtyps «Messen» stehen neben anderen Experimentiermaterialien jeweils zwei verschiedene Messinstrumente zur Verfügung, die sich in der Genauigkeit ihrer Skalen unterscheiden (vgl. Abb. 4).

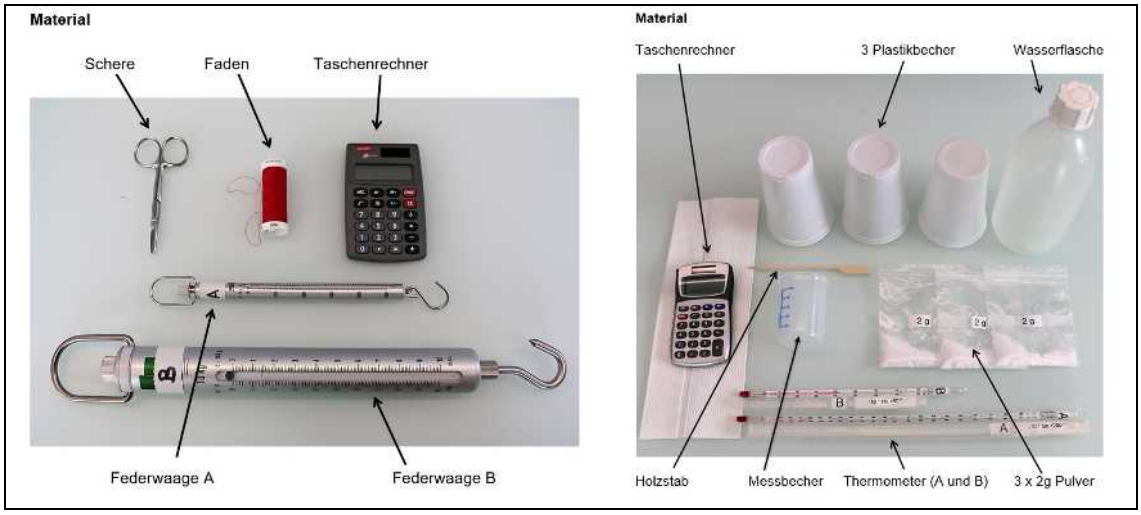


Abbildung 4: Das zur Verfügung stehende Experimentiermaterial am Beispiel der Fadenaufgabe (links) und der Pulveraufgabe (rechts). Fotos von Bonetti.

Beim Problemtyp «Messen» wurden sechs Aufgaben entwickelt (vgl. Tab. 6 und z. B. Gut et al., 2017; Metzger et al., 2014; Bonetti, in Vorbereitung). Da in der Schweiz der naturwissenschaftliche Unterricht in der Sekundarstufe I interdisziplinär stattfindet, sind auch die Aufgaben interdisziplinär angelegt: Zwar können die Kontexte der Aufgaben ansatzweisen einer Disziplin (Biologie, Chemie oder Physik) zugeordnet werden, über den gesamten Problemtyp hinweg sind die Disziplinen aber gleich stark vertreten[23]. Innerhalb eines Problemtyps wurde auf grösstmögliche Homogenität bei den schriftlichen Aufgabenstellungen und Antwortformaten geachtet, um kompetenzirrelevante Aufgabenanforderungen möglichst konstant zu halten. Um den Einfluss von kontextspezifischem Fachwissen zu minimieren, wurden hauptsächlich Alltagskontexte und alltagssprachliche Formulierungen verwendet (Metzger & Gut, 2017). Somit stehen bei den Aufgaben vor allem methodische Kompetenzen im Vordergrund, wie

[23] Zuordnung der Aufgabe zu einer naturwissenschaftlichen Disziplin: Ahorn und Bohne eher biologischer Bereich; Filzstift und Pulver eher chemischer Bereich; Faden und Münze eher physikalischer Bereich.

beispielsweise das Durchführen von Messwiederholungen. Ziel der Aufgaben des Problemtyps «Messen» ist es, eine gesuchte Grösse mit vorgegebenen Messinstrumenten möglichst genau zu messen. Hierfür sollen die Schülerinnen und Schüler das genauere Messinstrument erkennen und für ihre Messung wählen, Messungen wiederholen und eventuell mit einer Menge messen, um ein möglichst genaues Ergebnis zu erhalten (vgl. Tab. 6, intendierte Konzepte und Unterkapitel 2.3.1).

Tabelle 6: Aufgaben des Problemtyps «Messen» und intendierte Konzepte

Aufgabe	**Problemstellung / Kontext**	**Intendierte Konzepte**
Ahorn	Die SuS sollen herausfinden, wie lange ein Ahornsamen drehend braucht, um die Strecke einer Pulthöhe zu fallen.	<u>Messwiederholung:</u> Durch das Durchführen von Messwiederholungen und das Bilden eines Mittelwerts kann die Messgenauigkeit erhöht werden. <u>Mengenvergrösserung:</u> Das Messen mit einer (grossen) Menge kann die Messgenauigkeit erhöhen. Das Messergebnis der Menge kann auf die gesuchte Grösse zurückgerechnet werden. <u>Wahl Messinstrument:</u> Für eine genaue Messung sollte immer ein möglichst genaues Messinstrument gewählt werden.
Bohne	Die SuS sollen herausfinden, wie schwer eine getrocknete Bohne ist.	
Faden	Die SuS sollen herausfinden, bei welcher Belastung ein Nähfaden reisst.	
Filzstift	Die SuS sollen herausfinden, wie weit ein Filzstiftpunkt in Sprit in zwei Minuten wandert.	
Münze	Die SuS sollen herausfinden, wie viel Wasser eine 5 Rappen-Münze verdrängt.	
Pulver	Die SuS sollen herausfinden, wie sich die Temperatur von 50 ml Wasser verändert, wenn sie ein Tütchen Pulver hineingeben.	

Anmerkung: SuS steht für Schülerinnen und Schüler.

In Tabelle 6 werden die Problemstellungen und intendierten Konzepte der Aufgaben des Problemtyps «Messen» vorgestellt. Bei den Aufgaben sollen die Schülerinnen und Schüler über die Zusammenhänge zwischen der Anzahl Messungen, der Wahl des Messinstruments oder dem Messen mit einer Menge und der Messgenauigkeit nachdenken. Während die Messwiederholung mit Mittelwertbildung und die Wahl des genaueren Messinstruments als Beiträge zur Steigerung der Messgenauigkeit im Wesentlichen selbsterklärend sind, ist dies beim Messen mit einer Menge nicht unbedingt der Fall. Beispielsweise kann mit doppelter Pulthöhe (Ahornaufgabe), mit mehreren Bohnen auf einmal (Bohnenaufgabe), mit doppeltem Faden (Fadenaufgabe), mit mehreren Punkten

oder mehr als zwei Minuten (Filzstiftaufgabe), mit mehreren Münzen (Münzenaufgabe) oder mehreren Tütchen Pulver (Pulveraufgabe) aufs Mal gemessen und dieser Wert anschliessend auf die gesuchte Grösse zurückgerechnet werden. Da bei den Aufgaben nicht die fachinhaltlichen, sondern die methodischen Kompetenzen überprüft werden sollen, ist es in Ordnung, wenn bei der Filzstiftaufgabe angenommen wird, dass sich die gemessene Zeit und die zurückgelegte Strecke linear verhalten und mit mehr als zwei Minuten gemessen wird, bei der Pulveraufgabe der Temperaturunterschied und die Menge an Pulver linear zusammenhängen oder bei der Ahornaufgabe der Ahornsamen mit einer annähernd gleichbleibenden Geschwindigkeit zu Boden fällt. Durch die Strategie der Mengenvergrösserung kann die Genauigkeit der Messung erhöht werden, indem das Messergebnis besser auf der Skala des Messinstruments abgelesen werden kann (z. B. Bohnen-, Faden-, Filzstift-, Münzen- und Pulveraufgabe). Zudem können durch die Strategie der Mengenvergrösserung auch Unterschiede bei der zu messenden Grösse berücksichtigt werden (z. B. unterschiedliche Bohnengrössen oder minimale Abweichungen bei den Münzen). Hier gilt: Je grösser die Menge ist, mit der man misst, desto genauer wird das Messergebnis. Des Weiteren kann durch die Strategie der Mengenvergrösserung auch der Einfluss der Reaktionszeit minimiert werden, zum Beispiel bei der Ahornaufgabe die Reaktionszeit zum Bedienen der Stoppuhr. Das Messen mit einer Menge kann aber auch dazu dienen, mehrere Anhaltspunkte mit Hilfe einer Messung zu generieren: Wird bei der Filzstiftaufgabe mit mehreren Punkten gleichzeitig gemessen, so erhält man durch eine Messung mehrere Messwerte.

6.4.2 Schülerprotokolle

Während des Lösens der Problemstellungen mit Realexperimenten notierten die Schülerinnen und Schüler ihre Vorgehensweise, Messdaten, Ergebnisse, Überlegungen und Schlussfolgerungen in Schülerprotokollen (P). Die Schülerprotokolle sind durch offene und geschlossene Aufträge vorstrukturiert und umfassen beim Problemtyp «Messen» insgesamt sieben Seiten. Bei den offenen Aufträgen steht jeweils unterhalb des Auftrags freier Platz zur Verfügung: Hier können die Schülerinnen und Schüler ihre Vorgehensweisen beschreiben und ihre Antworten und Gedanken festhalten. Ein geschlossener Auftrag im Schülerprotokoll ist zum Beispiel, dass die Lernenden ankreuzen, mit welchem Messinstrument sie gemessen haben. Innerhalb eines Problemtyps sind die Aufträge standardisiert. Beim Problemtyp «Messen» sind folgende Aufträge im vorstrukturierten Schülerprotokoll enthalten:

- Beschreibe und skizziere, welche Messungen du gemacht hast.
- Was hast du herausgefunden?
  Vorgegebene Textbausteine zum Angeben der Ergebnisse:
  Ein Ahornsamen braucht …, um die Strecke einer Pulthöhe zu fallen.
  Das Gewicht einer einzelnen Bohne ist …
  Der Faden reisst bei einer Belastung von …
  Der Filzstiftpunkt wandert in 2 Minuten … weit.
  Die Verdrängung einer 5 Rappen-Münze beträgt …
  Die Temperatur des Wassers verändert sich um …, wenn man Pulver hineingibt.
- Erkläre, warum dein Resultat genau oder ungenau ist.
- Wie könntest du noch genauer messen? Mache Vorschläge und erkläre, wieso man so genauer messen kann.
- Mit welcher / welchem [Stoppuhr, Waage, Federwaage, Massstab, Messzylinder, Thermometer] hast du gemessen? Kreuze an.
- Kannst du mit beiden [Stoppuhren, Waagen, Federwaagen, Massstäben, Messzylindern, Thermometern] gleich genau messen? Begründe.
- Wie viel Male hast du gemessen, bis du dein Ergebnis hattest?
- Mit wie viel / vielen [Pulthöhen, Bohnen, Fäden, Filzstiftpunkten, Münzen, Pulver] hast du gemessen?
- Hast du etwas gerechnet, um dein Ergebnis zu erhalten? Erkläre.

Ausschnitte aus einem Schülerprotokoll sind in Unterkapitel 7.1.1 zu finden. Der Aufbau der Schülerprotokolle wird zudem detailliert im Dissertationsprojekt von Bonetti (Bonetti, in Vorbereitung) beschrieben.

6.4.3 Videoaufzeichnungen

Beim Bearbeiten der Aufgaben des Problemtyps «Messen» wurden die Schülerinnen und Schüler während der gesamten Zeitspanne (also während 18 Minuten) gefilmt (Videoaufnahmen, V). Da die Schülerinnen und Schüler die Aufgaben in Einzelarbeit bearbeiteten, beinhalten die Videos vor allem experimentelle Handlungen und keine verbalen Äusserungen. Die Kameraeinstellung (vgl. Abb. 5) erlaubt das Beobachten von experimentellen Vorgehensweisen (z. B. zur Problemstellung passende Vorgehensweise; Durchführen von Messwiederholungen; Messen mit einer Menge und Wahl des Messinstruments), während die Notizen der Lernenden in den Schülerprotokollen (z. B. Messdaten, Berechnungen, Ergebnisse und Begründungen) anhand der Videoaufnahmen nicht ersichtlich werden. Im Rahmen vorliegender Arbeit stellt dies aber kein Problem dar, da im

Zuge des Untersuchens von FF1 die Videos nicht isoliert, sondern zusammen mit den Schülerprotokollen ausgewertet wurden (also PV, vgl. Kapitel 5).

Abbildung 5: Beispiele für die Kameraeinstellung der Videoaufnahmen bei der Ahornaufgabe (links) und Bohnenaufgabe (rechts).

Abbildung 5 verdeutlicht, dass die Distanzen, aus welchen gefilmt wurden, je nach Aufgabenstellungen unterschiedlich waren: Während beispielsweise bei der Ahornaufgabe[24] aus grösserer Distanz gefilmt wurde, um auch das Messen mit doppelter Pulthöhe erfassen zu können, wurde zum Beispiel bei der Bohnenaufgabe[25] aus geringerer Distanz gefilmt, damit ungefähr mitgezählt werden konnte, mit wie vielen Bohnen die Schülerinnen und Schüler auf einmal messen.

6.4.4 Interviews

Nach der Bearbeitung der experimentellen Problemstellungen wurden die Schülerinnen und Schüler in Einzelinterviews zu den Aufgaben des Problemtyps «Messen» und ihrer Vorgehensweise, ihren Ergebnissen, ihren Schlussfolgerungen und Gedanken befragt (Interviews, I). Die Interviews wurden auf Video aufgezeichnet und dauerten zwischen 11 und 22 Minuten[26]. Für die Interviews standen den Lernenden die ausgefüllten Schülerprotokolle und die Experimentiermaterialien der Aufgabe als Erinnerungshilfen (Stimulus) zur Verfügung (vgl. Abb. 6).

[24] Ebenfalls aus grösserer Distanz gefilmt wurde bei der Fadenaufgabe, da manche Lernende den Faden z. B. am Stuhl- oder Tischbein befestigt haben.
[25] Ebenfalls aus geringerer Distanz gefilmt wurde bei der Filzstift-, Münzen- und Pulveraufgabe.
[26] Die Mehrheit der Interviews (ca. 68 % der Interviews) dauerten zwischen 13 und 18 Minuten.

Abbildung 6: Kameraeinstellung und Erinnerungshilfen (ausgefüllte Schülerprotokolle und Experimentiermaterialien) bei den Interviews am Beispiel der Bohnenaufgabe.

Die Interviews wurden in Anlehnung an Helfferich (2011) mit Hilfe eines Interviewleitfadens (vgl. Anhang, Teil A) geführt. Der Leitfaden umfasst offene Fragen (Prompts) im Bereich der intendierten Konzepte. Somit wurden den Lernenden Fragen zu Messstrategien (Messwiederholungen und Mengenvergrösserung) und zur Wahl des Messinstruments gestellt. Zudem wurden die Schülerinnen und Schüler aufgefordert, ihre Vorgehensweise zum Messen der gesuchten Grösse und ihre Ergebnisse zu erläutern. Falls in diesen Bereichen Notizen der Lernenden in den Schülerprotokollen vorhanden waren, wurde während des Interviews Bezug auf diese genommen (vgl. Abb. 7, oben). Ansonsten wurden die Lernenden aufgefordert, ihre Vorgehensweisen und Überlegungen ohne Notizen zu erläutern und erklären (vgl. Abb. 7, unten).

Wenn bei i08 eine Antwort vorhanden ist

Du schreibst, dass du mit _____Bohnen aufs Mal gemessen hast. Warum hast du genau mit ____ Bohnen aufs Mal gemessen?

Wenn mit mehreren Bohnen gemessen wurde: (i09)

Welches Ergebnis hast du dann für ____ Bohnen erhalten? Erkläre, wie du dann vorgegangen bist, um das Gewicht für eine Bohne zu erhalten.

Wenn bei i08 nicht steht, mit wie vielen Bohnen aufs Mal gemessen wurde

Mit wie vielen Bohnen hast du aufs Mal gemessen? Warum hast du genau _____ Bohnen aufs Mal genommen?

Wenn mit mehreren Bohnen gemessen wurde: (i09)

Welches Ergebnis hast du dann für ____ Bohnen erhalten? Erkläre, wie du dann vorgegangen bist, um das Gewicht für eine Bohne zu erhalten.

Abbildung 7: Ausschnitt aus dem Interviewleitfaden der Bohnenaufgabe im Bereich der Strategie Mengenvergrösserung. Verweise wie z. B. i08 bezeichnen den entsprechenden Indikator (bzw. die entsprechende Stelle) im vorstrukturierten Schülerprotokoll.

Neben den Fragen zu den intendierten Konzepten waren im Interviewleitfaden auch einige allgemeine Fragen integriert[27]. Diese wurden im Zuge der Auswertungen von FF1 und FF2 nicht genutzt, können aber zur weiterführenden Forschung dienen (vgl. Kapitel 9).

[27] Z. B. mussten die Lernenden anhand von vorgegebenen Messdaten erklären, wie sie zu einem Endresultat kommen könnten, in Anlehnung an Hellwig (2012) und Lubben & Millar (1996) (vgl. auch Anhang, Teil A).

7. Teilstudie I: Vergleich verschiedener Erhebungsmethoden am Beispiel von Aufgaben des Problemtyps «Messen»

Teilstudie I vergleicht verschiedene Erhebungsmethoden bei Tests mit Realexperimenten bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose (FF1) und prüft die damit verbundenen Hypothesen (H1 bis H5, vgl. Kapitel 5). Die im Rahmen von Teilstudie I genutzten Instrumente (Aufgaben des Problemtyps «Messen», Schülerprotokolle, Videoaufnahmen und Interviews) wurden bereits in Kapitel 6 beschrieben. In diesem Kapitel liegt somit der Fokus auf der Auswertung der Daten im Rahmen von Teilstudie I und den Ergebnissen.

7.1 Auswertung der Daten

In einem ersten Schritt der Datenauswertung wurde zunächst jede der drei Datenquellen – Schülerprotokolle (P), Videoaufzeichnungen (V) und Interviews (I) – getrennt ausgewertet (vgl. Unterkapitel 7.1.1). Da im Rahmen von FF1 untersucht wird, inwiefern *zusätzliche* Erhebungsmethoden (z. B. P vs. PV, P vs. PI) die Genauigkeit der Diagnostik erhöhen und inwiefern verschiedene Erhebungsmethoden (P, PV und PI) bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose an den gesetzten Benchmark (PVI) herankommen (vgl. Kapitel 5), wurden im Anschluss die Daten der Erhebungsmethoden kombiniert. Die Regeln zur Kombination werden in Unterkapitel 7.1.2 erläutert.

7.1.1 Auswertung der Schülerprotokolle, Videoaufnahmen und Interviews

Im Rahmen vorliegender Arbeit werden die Ergebnisse der Kompetenzdiagnose anhand verschiedener Erhebungsmethoden am Beispiel der Aufgaben des Problemtyps «Messen» verglichen. Dafür wurden für diese Aufgaben in einem ersten Schritt vergleichbare Kodiermanuale für die Kompetenzdiagnose anhand von Schülerprotokollen, Videoaufnahmen und Interviews entwickelt. Anhand der Kodiermanuale lassen sich unterschiedliche Ausprägungen experimenteller Kompetenzen erfassen, wobei die Auswertung produktorientiert ist: Es stehen das Verwenden von Messstrategien (z. B. Messwiederholungen und Mengenvergrösserung), die Wahl des Messinstruments und die Begründung hierzu, die Ergebnisse sowie Schlussfolgerungen im Fokus (im Vergleich zu einer prozessorientierten Auswertung, bei welcher auch der zeitliche Verlauf während des Experimentierens eine Rolle spielt, vgl. Kapitel 3). Zuerst wurden die Kodiermanuale der Schülerprotokolle erstellt, welche auch im Rahmen der Gesamtvalidierungsstudie (vgl. Unterkapitel 6.1) bei der Auswertung der Aufgaben des Problemtyps «Messen» eingesetzt wurden. Die Kodiermanuale der Schülerprotokolle dienten

als Ausgangspunkt für die Entwicklung der Manuale der Videoaufnahmen und Interviews. Im Folgenden wird der Aufbau der Kodiermanuale am Beispiel des Manuals der Schülerprotokolle beschrieben. Anschliessend wird die Vorgehensweise bei der Entwicklung der Manuale für die Videoaufnahmen und Interviews erläutert. Zum Schluss wird auf die Kodierung der Schülerprotokolle, Videoaufnahmen und Interviews eingegangen, zum Beispiel auf den Zeitraum der Kodierung und die Doppelkodierung.

Aufbau der Kodiermanuale

Für jede Aufgabe des Problemtyps «Messen» wurde ein Kodiermanual entwickelt, wobei die Manuale der Aufgaben analog aufgebaut sind, das heisst die gleichen Indikatoren beinhalten. Die Kodiermanuale umfassen zu den verschiedenen Qualitätsstandards (vgl. Unterkapitel 2.3.2) jeweils zwischen drei und neun Indikatoren. Um eine möglichst objektive Kodierung zu ermöglichen, werden die Indikatoren durch Kodiervorschriften und Explikationen respektive Ankerbeispiele unterstützt (vgl. Tab. 7). Bei den meisten Indikatoren erhalten die Schülerinnen und Schüler bei Erfüllung des Indikators einen Punkt[28]. Anhand dieser Indikatoren und der somit erreichten Punkteanzahl werden unterschiedliche Ausprägungen experimenteller Kompetenzen seitens der Schülerinnen und Schüler diagnostiziert. In Tabelle 7 ist ein Beispiel der Bepunktung eines Indikators der Fadenaufgabe dargestellt.

Tabelle 7: Beispiel der Bepunktung eines Indikators der Fadenaufgabe aus dem Bereich Messwiederholungen (Qualitätsstandard 3, vgl. Unterkapitel 2.3.2).

Messwiederholung als Strategie, um ein möglichst genaues Ergebnis zu erhalten		
m3.1	**1P**	*Notwendigen Bedingungen*: a. Es sind Daten zu einer Messwiederholung vorhanden. Gilt auch, wenn klar wird, dass nicht auf exakt die gleiche Weise gemessen wurde. b. Es wird ein Wert aus der Messreihe als Resultat berechnet / ausgewählt. c. Die ausgewählte Messreihe für die Berechnung / Auswahl des Resultats wurde mit demselben Messinstrument erstellt. (Wenn klar wird, dass für das Resultat unterschiedliche Messinstrumente verwendet wurden, ist dies nicht als Messwiederholung zu werten.)

[28] Im Manual gab es auch einige Indikatoren, die nicht bepunktet, sondern anhand von Kategorien (Codes) ausgewertet wurden. Diese Kodierung wurde v.a. im Rahmen der Gesamtvalidierungsstudie (vgl. Unterkapitel 6.1) genutzt, um mehr über die Vorgehensweisen der Schülerinnen und Schüler zu erfahren. Z. B. wurde im Bereich Messwiederholung anhand von Codes erfasst, wie die Lernenden mit den Daten der Messwiederholung weiterarbeiten (Mittelwertbildung; Angabe eines Spektrums als Resultat; Wahl des bestätigten Werts; etc.). Auch Qualitätsstandard 5 wurde anhand von Codes ausgewertet (vgl. Unterkapitel 2.3.2).

		Explikationen / Ankerbeispiele: a. z. B. auch gültig: Einmal Faden am Tischbein befestigt und einmal Faden von Hand gezogen. b. Gilt, wenn z. B. «3 Messungen, Faden reisst zw. 900-1000 g» als Resultat angegeben wird. c. Gilt nicht, wenn z. B. Federwaage A: 1100 g, Federwaage B: 1200 g

Anmerkung: m3.1 ist die Bezeichnung für den entsprechenden Indikator.

In Abbildung 8 sind Ausschnitte aus einem Protokoll eines Schülers ersichtlich, um die Kodierung zu illustrieren. Beim Betrachten der Ausschnitte wird deutlich, dass der Schüler den Indikator m3.1 aus Tabelle 7 erfüllt hat (somit 1 Punkt), da alle drei notwendigen Bedingungen erfüllt sind (a. Daten zu Messwiederholungen vorhanden; b. ein Wert aus Messreihe als Resultat berechnet und c. Messreihe mit dem gleichen Messinstrument erstellt).

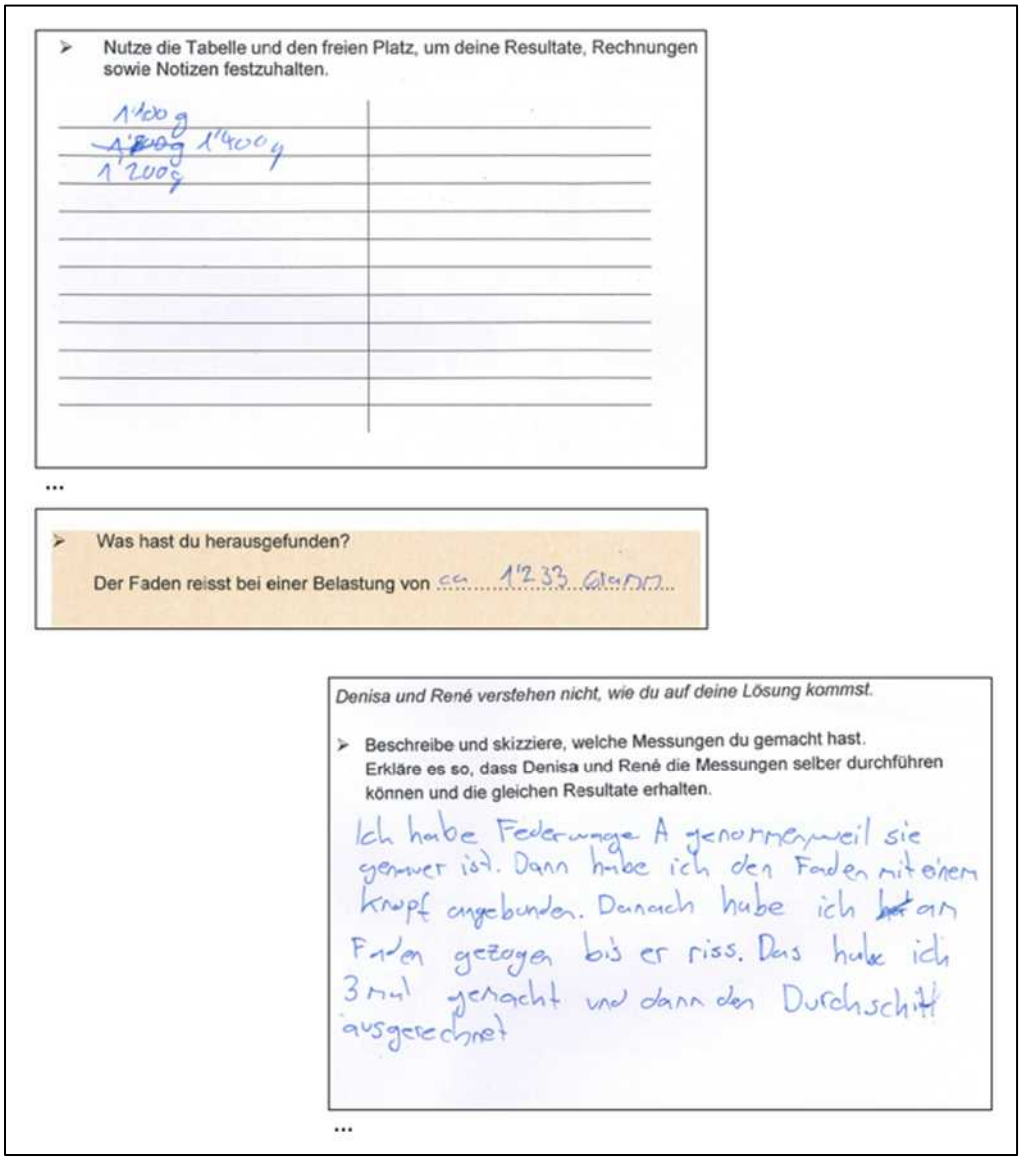


Abbildung 8: Ausschnitte aus dem Protokoll eines Schülers bei der Fadenaufgabe.

Erstellen der Kodiermanuale für die Videoaufnahmen und Interviews

Um die Daten der Erhebungsmethoden vergleichen zu können, wurden analog zum Kodiermanual der Schülerprotokolle, welches als Ausgangspunkt diente, die Manuale zu den Videoaufnahmen und Interviews entwickelt. Die Manuale für die Schülerprotokolle, Videoaufnahmen und Interviews beinhalten die gleichen Indikatoren, wobei die Indikatoren formal für die jeweilige Erhebungsmethode angepasst wurden. Die Anpassung der Indikatoren für die jeweilige Erhebungsmethode wird in Tabelle 8 anhand eines Beispiels illustriert.

Tabelle 8: Anpassen der Indikatoren für die jeweilige Erhebungsmethode am Beispiel eines Indikators aus dem Bereich 'adäquate Vorgehensweise' (Qualitätsstandard 1, vgl. Unterkapitel 2.3.2)

Schülerprotokolle	**Videoaufnahmen**	**Interviews**
m1.2 (1P): Einzelmessung Die protokollierte Vogehensweise passt zur Problemstellung, d.h. ist adäquat zum Messen der gesuchten Grösse.	**m1.2 (1P): Einzelmessung** Die im Video ersichtliche Vogehensweise passt zur Problemstellung, d.h. ist adäquat zum Messen der gesuchten Grösse.	**m1.2 (1P): Einzelmessung** Die im Interview beschriebene Vogehensweise passt zur Problemstellung, d.h. ist adäquat zum Messen der gesuchten Grösse.

Anmerkung: m1.2 ist die Bezeichnung für den entsprechenden Indikator.

Kodierung der Schülerprotokolle, Videoaufnahmen und Interviews anhand der Kodiermanuale

Die Schülerprotokolle (P) wurden im Rahmen der Gesamtvalidierungsstudie (vgl. Unterkapitel 6.1) kodiert. Die Kodierung der Schülerprotokolle fand im Januar und Februar 2018 durch studentische Hilfskräfte statt, die vorgängig eingehend geschult und zudem beim Kodierprozess begleitet wurden. Um zu entscheiden, ob ein Indikator erfüllt ist, wurden alle Notizen im Schülerprotokoll berücksichtigt, das heisst, dass die Antwort nicht zwingend an der richtigen Stelle im Protokoll notiert sein musste. Um eine objektive Kodierung zu gewährleisten, wurden 15 % der Schülerprotokolle doppelt kodiert. Vor Beginn der Doppelkodierung wurde zunächst, wie von Wirtz und Caspar (2002) empfohlen, eine Kodier-Schulung durchgeführt. Hierbei wurden die Inhalte des Kodiermanuals besprochen und die Kodierung anhand einiger Beispiele illustriert. In Anlehnung an Wirtz und Caspar (2002) wurden zur Einschätzung der Beurteilerübereinstimmung Cohens Kappa (Cohens k) und die prozentuale Übereinstimmung (pÜ) verwendet. Im Rahmen der doppelten Kodierung der Schülerprotokolle konnte dabei

eine zufriedenstellende Beurteilerübereinstimmung erreicht werden (Cohens $k \geq .61$; pÜ ≥ 81 %)[29].

Die Videoaufnahmen von den Schülerinnen und Schülern während des Experimentierens (V) und die Interviews zu den Aufgaben (I) wurden im Rahmen vorliegender Studie von zwei Kodiererinnen (Autorin dieser Arbeit und eine Projektmitarbeiterin) ausgewertet. Die Kodierung der Videos und Interviews fand von Mai bis Juli 2018 statt. Zur Kodierung der Interviews wurden die Videoaufzeichnungen der Interviews betrachtet. Das bedeutet, dass im Rahmen vorliegender Studie, unter anderem aus testökonomischen Gründen, auf eine Transkribierung der Interviews verzichtet wurde. Für die Kodierung der Videos und Interviews wurde die gesamte Zeitspanne, die Informationen über einen entsprechenden Indikator liefern kann, betrachtet und danach der Indikator beurteilt. Wenn ein Indikator nicht beurteilt werden konnte, dann wurde ein Missing gesetzt. Bei den Videos (V) konnten einige Indikatoren nicht beurteilt werden, da die Aufnahmen primär experimentelle Handlungen zeigen (vgl. Unterkapitel 6.4.3). Ein Beispiel für einen Indikator, der anhand der Videos nicht beurteilt werden kann, ist Indikator m3.1: Anhand der Videos wird nicht ersichtlich, ob ein Wert als Resultat ausgewählt oder berechnet wurde (vgl. Tab. 7, notwendige Bedingung b). Die Missings spielen im Rahmen der Auswertungen zu FF1 jedoch keine Rolle, da hierfür die Videos (und auch Interviews) nicht isoliert, sondern zusammen mit den Schülerprotokollen betrachtet wurden (PV, PI bzw. PVI, vgl. Kapitel 5). Analog zum Vorgehen bei den Schülerprotokollen fand vor Beginn der Kodierung der Videos und Interviews zunächst eine Kodier-Schulung statt. Daraufhin wurden 30 % der Videos und Interviews doppelt kodiert, wobei eine zufriedenstellende Übereinstimmung erreicht werden konnte (Videos: Cohens $k \geq .77$, pÜ ≥ 91 %; Interviews: Cohens $k \geq .70$, pÜ ≥ 86 %; wobei die Übereinstimmung für alle Indikatoren bestimmt wurde (vgl. auch Fussnote 29)).

7.1.2 Kombination der Daten der Erhebungsmethoden

Im Rahmen des Untersuchens von FF1 wurden die Daten der Schülerprotokolle (P), der Videoaufnahmen während des Experimentierens (V) und der Interviews (I) kombiniert, womit die Genauigkeit der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI (gesetzter Benchmark, vgl. Kapitel 5) verglichen werden können. Für die Kombination der Daten der Erhebungsmethoden wurden die

[29] Gemäss Landis & Koch (1977) gilt ein Wert ab 0.6 für Cohens *k* als substantiell und somit zufriedenstellend. Die Beurteilerübereinstimmung wurde für die bepunkteten Indikatoren und die Indikatoren, die anhand von Kategorien (Codes) ausgewertet wurden (vgl. Fussnote 28), bestimmt. Bei den Schülerprotokollen betrug hierbei der Wert für Cohens *k* jeweils mindestens 0.61.

bepunkteten Indikatoren betrachtet, welche auch in das Ergebnis der Kompetenzdiagnose miteinfliessen, und es galt die Regel, dass, sobald ein Indikator in einem Zugang gezeigt und somit als erfüllt beurteilt wurde, dieser in der Kombination der Daten als erfüllt gilt. Die Vorgehensweise zur Kombination der Daten der Erhebungsmethoden wird in der Folge anhand von Beispielen illustriert.

Beispiel 1:
Ein Indikator ist anhand der Schülerprotokolle (P) erfüllt. In den Videoaufnahmen (V) wird nicht ersichtlich, ob der Indikator erfüllt ist. Folglich gilt der Indikator in der Kombination PV als erfüllt.
Beispiel: Beim Betrachten der Ausschnitte aus dem Schülerprotokoll (vgl. Abb. 8) konnte festgestellt werden, dass Indikator m3.1 anhand des Schülerprotokolls als erfüllt gilt. In der Videoaufnahme wird nicht ersichtlich, ob Indikator m3.1 erfüllt ist, da anhand der Aufnahme nicht beurteilt werden kann, ob ein Wert aus der Messreihe als Resultat ausgewählt oder berechnet wurde (vgl. Tab. 7, notwendige Bedingung b). In der Kombination PV ist Indikator m3.1 jedoch erfüllt, da dieser in einem Zugang als erfüllt beurteilt wurde.

Beispiel 2:
Ein Indikator ist anhand der Schülerprotokolle (P) nicht erfüllt, da hierzu im Schülerprotokoll nichts notiert wurde. In den Videoaufnahmen (V) wird ersichtlich, dass der Indikator erfüllt ist. Folglich gilt der Indikator in der Kombination PV als erfüllt.
Beispiel: Indikator m1.2 (adäquate Vorgehensweise, vgl. Tab. 8) ist anhand eines Schülerprotokolls nicht erfüllt, da keine Vorgehensweise protokolliert wurde. In der Videoaufnahme wird jedoch ersichtlich, dass die durchgeführte Vorgehensweise adäquat zum Messen der gesuchten Grösse ist (vgl. Abb. 9) und somit wird Indikator m1.2 anhand der Videoaufnahme als erfüllt beurteilt. Damit ist Indikator m1.2 in der Kombination PV erfüllt.

Abbildung 9: In den Videoaufnahmen ersichtliche adäquate Vorgehensweise zur Bestimmung, bei welcher Belastung ein Nähfaden reisst (Fadenaufgabe).

Beispiel 3:
Anhand des Schülerprotokolls (P) ist ein Indikator nicht erfüllt, da in diesem Bereich nichts notiert wurde. In der Videoaufnahme (V) wird der Indikator nicht ersichtlich, da es sich vorwiegend um kognitive Prozesse handelt. Anhand der Erläuterungen im Interview (I) wird jedoch deutlich, dass der Indikator erfüllt ist. Somit ist der Indikator in der Kombination PVI erfüllt.
Beispiel: Im Schülerprotokoll sind keine Daten zu Messwiederholungen und keine Rechnung oder Ähnliches notiert, sondern es wird lediglich ein Schlussresultat angegeben (angegebenes Schlussresultat, Beispiel Ahornaufgabe: *«Circa 1 Sekunde oder eher weniger»*). Anhand des Schülerprotokolls wird somit nicht ersichtlich, ob Messwiederholungen durchgeführt wurden und Indikator m3.1 (vgl. Tab. 7) wird als nicht erfüllt beurteilt. In der Videoaufnahme wird zwar ersichtlich, dass mehrmals gemessen wurde, Indikator m3.1 kann aber nicht beurteilt werden, da die notwendige Bedingung b (vgl. Tab. 7, ein Wert als Resultat berechnet / ausgewählt) anhand der Aufnahme nicht beurteilt werden kann. Anhand der Erläuterungen im Interview (Erläuterung der Schülerin: *«Ich habe mehrmals mit Stoppuhr B gemessen und habe dann gesehen, dass es immer 1 Sekunde oder eher weniger war»*) wurde Indikator m3.1 als erfüllt beurteilt, da klar wurde, dass die notwendigen Bedingungen a bis c von Indikator m3.1 (vgl. Tab. 7; a: Daten zu Messwiederholungen vorhanden; b: ein Wert als Resultat berechnet / ausgewählt; c: Messungen mit dem gleichen Messinstrument durchgeführt) erfüllt wurden. Somit ist Indikator m3.1 in der Kombination PVI erfüllt, da dieser in einem Zugang als erfüllt beurteilt wurde.

Sollten sich die Daten bei den Erhebungsmethoden ‘widersprechen’, dann wurde für die Kombination die bessere Lösung berücksichtigt. ‘Widersprüche’ können dann vorkommen, wenn ein Indikator grundsätzlich in mehreren Erhebungsmethoden ersichtlich, aber unterschiedlich beurteilt wird. Ein Beispiel für einen solchen ‘Widerspruch’ ist, dass ein Indikator anhand des Schülerprotokolls als erfüllt, aber anhand des Interviews als nicht erfüllt beurteilt wurde. Dies kam beispielsweise bei der Pulveraufgabe im Bereich der Wahl des Messinstruments vor (Qualitätsstandard 4, Indikator m4.1: «Es wird das genauere Messinstrument erkannt und die Begründung hierfür ist korrekt, Verweis auf die genauere Skala»). Anhand des Schülerprotokolls ist der Indikator erfüllt, da in den Notizen auf die genauere Skala von Thermometer B verwiesen wurde (1 Punkt). Im Interview wurde jedoch argumentiert, dass Thermometer B gewählt wurde, da dieses kürzer ist (vgl. auch Unterkapitel 6.4.1, Abb. 4) und somit der Becher mit Wasser nicht umkippt. Im Interview fehlte ein Verweis auf die genauere Skala von Thermometer B, wodurch der Indikator anhand des Interviews als nicht erfüllt beurteilt

wurde (0 Punkte). In solchen Fällen, bei welchen sich die Beurteilung eines Indikators anhand verschiedener Datenquellen unterscheiden (bzw. 'widersprechen'), wurde die bessere Lösung für die Kombination der Daten berücksichtigt. Insgesamt kamen jedoch 'widersprüchliche' Fälle nur wenig vor[30].

7.2 Ergebnisse

Um zunächst zu prüfen, ob die Erhebungsmethoden zur Diagnose der experimentellen Kompetenzen in der vorgegebenen Stichprobe geeignet sind und die Ergebnisse somit sinnvoll interpretiert werden können, wurde die Reliabilität der einzelnen Erhebungsmethoden respektive deren Kombinationen (P, PV, PI und PVI) mit Hilfe von Cronbachs Alpha geprüft. Die Prüfung ergab Cronbachs Alpha Werte zwischen .64 und .70 und fiel somit nach Massstäben für Cronbachs Alpha akzeptabel bis gut aus. P, PV, PI und PVI sind somit für die Diagnose experimenteller Kompetenzen in der gegebenen Stichprobe geeignet.
In den folgenden Ausführungen werden die Ergebnisse auf Ebene der Stichprobe (Unterkapitel 7.2.1) und auf Individualebene (Unterkapitel 7.2.2) berichtet. Beim Betrachten der Ergebnisse auf Ebene der Stichprobe stehen *nicht* die Ergebnisse der Kompetenzdiagnose einzelner Schülerinnen respektive Schüler im Fokus, sondern viel eher wird analysiert, inwiefern die Methoden beim Betrachten der gesamten Stichprobe zu ähnlichen Ergebnissen führen und somit auf Ebene der Stichprobe austauschbar sind. Im Vergleich hierzu wird beim Betrachten der Ergebnisse auf Individualebene berichtet, inwiefern auf der Ebene von einzelnen Schülerinnen und Schülern die Erhebungsmethoden zu ähnlichen Ergebnissen der Kompetenzdiagnose gelangen und somit auf individueller Ebene austauschbar sind. Das Betrachten der Ergebnisse auf beiden Ebenen ist sinnvoll, da es sein kann, dass die Erhebungsmethoden auf Individualebene nicht substituierbar sind, während diese auf Ebene der Stichprobe zu sehr ähnlichen Ergebnissen führen. Dies würde für eine Substituierbarkeit der Erhebungsmethoden zum Beispiel bei large-scale Assessments sprechen, während diese beim Einsatz für Zwecke der Individualdiagnostik nicht ausgetauscht werden können.

[30] Insgesamt wurden 1512 Indikatoren zur Diagnose der experimentellen Kompetenzen beurteilt (6 Aufgaben des Problemtyps «Messen» × 14 bepunktete Indikatoren pro Aufgabe × 18 Lernende pro Aufgabe). Hier kam es insgesamt zu 41 'widersprüchlichen' Kodierungen (≙ 2.7 % der Fälle).

7.2.1 Vergleich der Ergebnisse der Kompetenzdiagnose auf Ebene der Stichprobe

Zum Vergleich der Ergebnisse der Diagnose experimenteller Kompetenzen auf Ebene der Stichprobe wird zuerst die Verteilung der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI graphisch betrachtet. Anschliessend werden die Mittelwerte der Ergebnisse der Kompetenzdiagnose durch die verschiedenen Erhebungsmethoden auf Unterschiede untersucht.

Abbildung 10 zeigt die Verteilung der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI. Insgesamt konnten 0 bis 15 Punkte pro Aufgabe des Problemtyps «Messen» erreicht werden und total liegen 108 Ergebnisse der Kompetenzdiagnose für die jeweiligen Erhebungsmethoden vor (vgl. Unterkapitel 6.2). Beim Betrachten von Abbildung 10 kann festgestellt werden, dass die Diagnose experimenteller Kompetenzen durch PI und PVI auf Ebene der Stichprobe zu sehr ähnlichen Ergebnissen gelangt. Des Weiteren kann festgestellt werden, dass die Ergebnisse der Kompetenzdiagnose durch PI und PVI tendenziell höher sind als diejenigen durch P und PV.

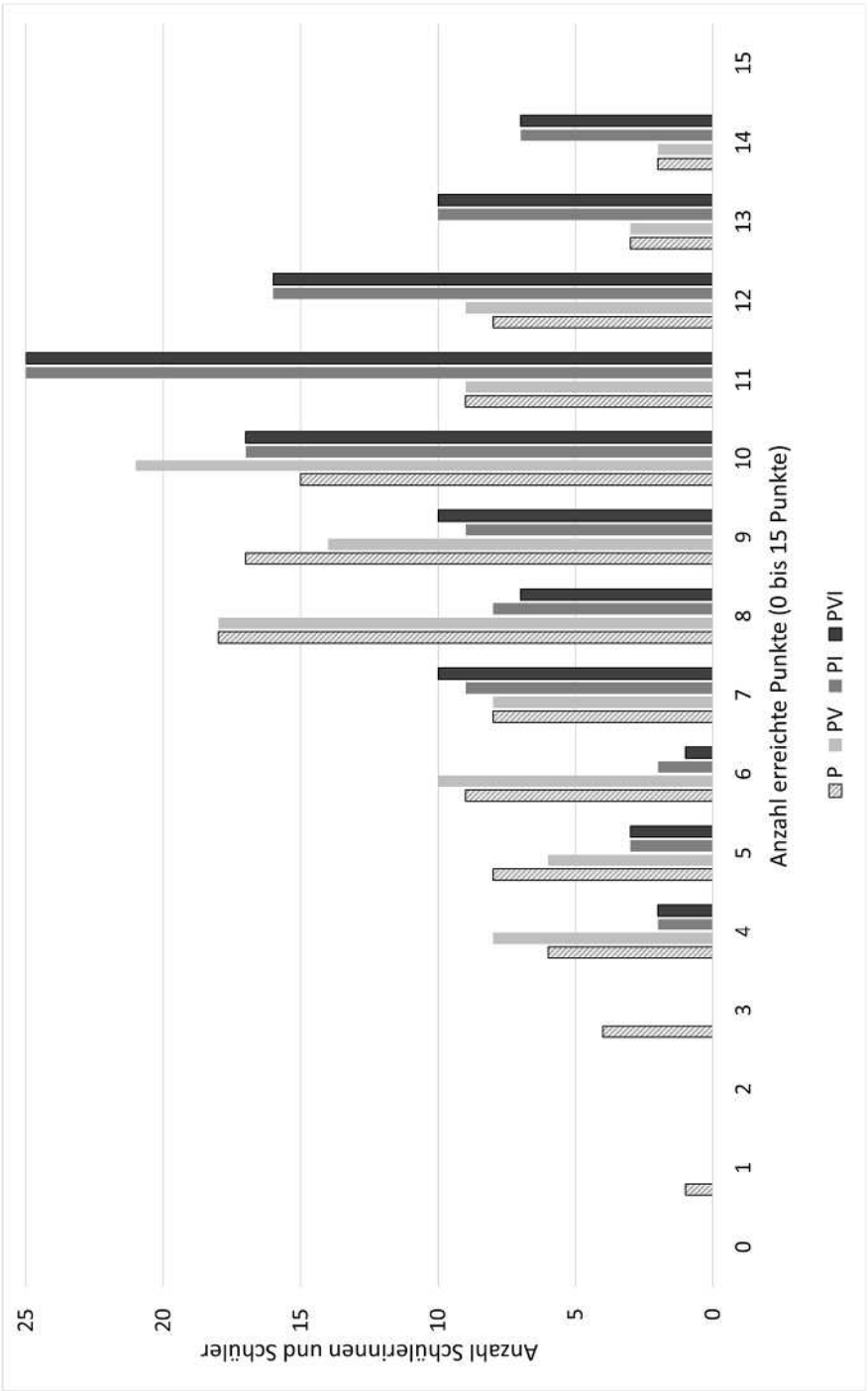


Abbildung 10: Verteilung der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI bei den Aufgaben des Problemtyps «Messen» (0 bis 15 Punkte möglich) durch die verschiedenen Erhebungsmethoden. Pro Erhebungsmethode: N = 108.

In Tabelle 9 werden die Mittelwerte von P, PV, PI und PVI sowie die Ergebnisse der Prüfung von Mittelwertsunterschieden aufgeführt. Zur Prüfung der Mittelwertsunterschiede wurde ein t-Test für abhängige Stichproben verwendet. Es handelt sich um eine abhängige Stichprobe, da die Daten der Erhebungsmethoden von denselben Schülerinnen und Schülern stammen und sich somit gegenseitig beeinflussen. Als Mass für die Effektstärke wird Cohens *d* angegeben, wobei die Werte gemäss Cohen (1988) wie folgt zu interpretieren sind: Kleiner Effekt $|d| \geq 0.2$, mittlerer Effekt $|d| \geq 0.5$ und grosser Effekt $|d| \geq 0.8$. Damit sich die Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI auf Ebene der Stichprobe *nicht* bedeutsam unterscheiden und somit austauschbar sind, wurde im Rahmen vorliegender Arbeit festgelegt, dass entweder keine signifikanten Unterschiede vorliegen dürfen oder die Unterschiede zwar signifikant aber höchstens als kleine Effekte ($|d| < 0.5$) einzustufen sind. Dies wurde so festgelegt, da im Rahmen der Auswertung eher geringe Unterschiede zwischen den Erhebungsmethoden erwartet werden, denn: (1) Bei den Erhebungsmethoden fand eine Auseinandersetzung mit den gleichen Teilaufträgen[31] statt und (2) die experimentellen Kompetenzen wurden mit sehr ähnlichen Kodiermanualen (vgl. Unterkapitel 7.1.1) diagnostiziert.

Tabelle 9: Mittelwerte und Mittelwertsunterschiede der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI bei den Aufgaben des Problemtyps «Messen» (0 bis 15 Punkte möglich).

	P	PV	PI	PVI	*N*
Mittelwerte (*M*)	8.30	8.63	10.24	10.28	108
Standardabweichungen (*SD*)	2.71	2.48	2.34	2.32	
Prüfen der Mittelwertsunterschiede	P vs. PV	t = -5.40	$p \leq .001$	d = -0.13	
	P vs. PI	t = -11.92	$p \leq .001$	d = -0.77	
	P vs. PVI	t = -12.14	$p \leq .001$	d = -0.79	
	PV vs. PVI	t = -12.09	$p \leq .001$	d = -0.69	
	PI vs. PVI	t = -1.42	p = .16 (n.s.)	-	

Anmerkung: t-Test für abhängige Stichproben und Cohens d als Mass für die Effektstärken.

[31] Die Aufträge der vorstrukturierten Schülerprotokolle (vgl. Unterkapitel 6.4.2) sind die Grundlage für die Auseinandersetzung mit den experimentellen Problemstellungen und bilden auch die Basis für die Videoaufnahmen und Interviews (in den Interviews wurden u.a. Nachfragen zu den Aufträgen der Schülerprotokolle gestellt, vgl. Unterkapitel 6.4.4).

Beim Betrachten der Mittelwerte in Tabelle 9 kann festgestellt werden, dass die experimentellen Kompetenzen der Schülerinnen und Schüler durch PI und PVI durchschnittlich um circa 1.5 bis 2 Punkte höher diagnostiziert werden als durch P und PV. Zudem zeigt Tabelle 9, dass zwischen PI und PVI kein signifikanter Mittelwertsunterschied vorliegt und somit die Methoden auf Ebene der Stichprobe zu einem sehr ähnlichen Ergebnis der Kompetenzdiagnose führen. Des Weiteren wird in Tabelle 9 ersichtlich, dass sich P und PV zwar signifikant unterscheiden, der Effekt aber als sehr gering ($|d| < 0.2$) und somit nicht bedeutsam einzustufen ist. Somit scheinen auch P und PV auf Ebene der Stichprobe zu einem ähnlichen Ergebnis der Kompetenzdiagnose zu gelangen. Im Vergleich dazu sind die Mittelwertsunterschiede zwischen P und PI, P und PVI sowie PV und PVI signifikant und bedeutsam (mittlere Effekte, $|d| \geq 0.5$), womit sich die Ergebnisse der Kompetenzdiagnosen durch P und PI, P und PVI sowie PV und PVI auf Ebene der Stichprobe beachtlich zu unterscheiden scheinen.

In Tabelle 10 werden die Mittelwerte und Mittelwertsunterschiede auf Ebene einzelner Qualitätsstandards (QS, vgl. Unterkapitel 2.3.2) betrachtet. Dadurch kann festgestellt werden, ob die in Tabelle 9 beobachteten bedeutsamen Mittelwertsunterschiede womöglich hauptsächlich bei einzelnen Qualitätsstandards entstehen.

Tabelle 10: Mittelwerte und Mittelwertsunterschiede der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI auf Ebene einzelner Qualitätsstandards beim Problemtyp «Messen»; N = 108.

<table>
<tr><th></th><th>Mittelwerte (M) und Standardabweichungen (SD)</th><th>Prüfen der Mittelwertsunterschiede (t-Test für abhängige Stichproben und Cohens d als Mass für die Effektstärken)</th></tr>
<tr><td>QS 1
(adäquate Vorgehensweise)</td><td>
<table>
<tr><th></th><th>M</th><th>SD</th></tr>
<tr><td>P</td><td>2.56</td><td>0.73</td></tr>
<tr><td>PV</td><td>2.71</td><td>0.51</td></tr>
<tr><td>PI</td><td>2.81</td><td>0.46</td></tr>
<tr><td>PVI</td><td>2.83</td><td>0.42</td></tr>
</table>
</td><td>
<table>
<tr><td>P vs. PV</td><td>t = -3.75 p ≤ .001 d = -0.24</td></tr>
<tr><td>P vs. PI</td><td>t = -4.17 p ≤ .001 d = -0.41</td></tr>
<tr><td>P vs. PVI</td><td>t = -4.42 p ≤ .001 d = -0.45</td></tr>
<tr><td>PV vs. PVI</td><td>t = -3.53 p ≤ .001 d = -0.26</td></tr>
<tr><td>PI vs. PVI</td><td>t = -1.42 p = .16 (n.s.) -</td></tr>
</table>
</td></tr>
</table>

QS 2 (eindeutiges Resultat mit korrekter Einheit)

	M	*SD*
P	2.29	0.98
PV	2.29	0.98
PI	2.48	0.78
PVI	2.48	0.78

P vs. PV	t-Test kann nicht ausgeführt werden, da der Standardfehler der Differenz gleich 0 ist.
P vs. PI	t = -3.76 $p \leq .001$ d = -0.22
P vs. PVI	t = -3.76 $p \leq .001$ d = -0.22
PV vs. PVI	t = -3.76 $p \leq .001$ d = -0.22
PI vs. PVI	t-Test kann nicht ausgeführt werden, da der Standardfehler der Differenz gleich 0 ist.

QS 3 (Messstrategie: Messwiederholung und Mengenvergrösserung)

	M	*SD*
P	1.94	1.40
PV	2.05	1.37
PI	2.85	1.39
PVI	2.85	1.39

P vs. PV	t = -3.66 $p \leq .001$ d = -0.08
P vs. PI	t = -9.30 $p \leq .001$ **d = -0.65**
P vs. PVI	t = -9.30 $p \leq .001$ **d = -0.65**
PV vs. PVI	t = -8.98 $p \leq .001$ **d = -0.58**
PI vs. PVI	t-Test kann nicht ausgeführt werden, da der Standardfehler der Differenz gleich 0 ist.

QS 4 (Messinstrument)

	M	*SD*
P	1.52	0.90
PV	1.58	0.82
PI	2.09	0.68
PVI	2.09	0.68

P vs. PV	t = -2.72 $p \leq .01$ d = -0.07
P vs. PI	t = -8.37 $p \leq .001$ **d = -0.72**
P vs. PVI	t = -8.37 $p \leq .001$ **d = -0.72**
PV vs. PVI	t = -8.35 $p \leq .001$ **d = -0.68**
PI vs. PVI	t-Test kann nicht ausgeführt werden, da der Standardfehler der Differenz gleich 0 ist.

Anmerkung: Fett hervorgehoben sind signifikante und bedeutsame Mittelwertsunterschiede ($|d| \geq 0.5$). Es wurden diejenigen QS betrachtet, die im Rahmen der Kompetenzdiagnose bepunktet wurden, also QS 1 bis QS 4. Bei QS 1, QS 2 und QS 4 konnten jeweils 0 bis 3 Punkte, bei QS 3 0 bis 6 Punkte erreicht werden.

Tabelle 10 zeigt, dass die bedeutsamen Mittelwertsunterschiede von Tabelle 9 womöglich hauptsächlich bei Qualitätsstandard 3 und 4 entstehen. Bei QS 3 und QS 4 liegen zwischen P und PI, P und PVI sowie PV und PVI signifikante und bedeutsame Mittelwertsunterschiede vor ($|d| \geq 0.5$), womit sich bei QS 3 und QS 4 die Ergebnisse der Kompetenzdiagnosen durch P und PI, P und PVI sowie PV und PVI auf Ebene der Stichprobe beachtlich zu unterscheiden scheinen.

In Tabelle 11 wird untersucht, ob bei QS 3 und QS 4 womöglich einzelne Indikatoren identifiziert werden können, die hauptsächlich zu den Unterschieden führen. Falls einzelne Indikatoren identifiziert werden können, dann zeigt dies auf, dass an diesen Stellen die Schülerprotokolle wahrscheinlich nicht für eine möglichst genaue Diagnose der experimentellen Kompetenzen ausreichen und allenfalls weitere Erhebungsmethoden, zum Beispiel Interviews, für ein genaueres Ergebnis der Kompetenzdiagnose nötig wären. Um Unterschiede auf Ebene einzelner Indikatoren zu lokalisieren, werden in Tabelle 11 die Mittelwerte der Indikatoren von QS 3 und QS 4 aufgeführt und deskriptiv betrachtet.

Tabelle 11: Mittelwerte der Ergebnisse der Kompetenzdiagnose bei den Indikatoren von Qualitätsstandard 3 und 4 des Problemtyps «Messen»; N = 108.

	Indikator	**Beschreibung**	**Erhebungsmethode**			
			P	**PV**	**PI**	**PVI**
QS 3 (Messstrategie)	m3.1	Messwiederholungen (MW) werden durchgeführt und ein Wert als Resultat ausgewählt / berechnet.	***M* = 0.55** *SD* = 0.50	***M* = 0.55** *SD* = 0.50	***M* = 0.84** *SD* = 0.37	***M* = 0.84** *SD* = 0.37
	m3.3	Aus den Daten der MW wird das arithmetische Mittel berechnet.	*M* = 0.41 *SD* = 0.49	*M* = 0.41 *SD* = 0.49	*M* = 0.52 *SD* = 0.50	*M* = 0.52 *SD* = 0.50
	m3.4	MW werden als Lösungsvorschlag zur Steigerung der Messgenauigkeit angegeben.	*M* = 0.15 *SD* = 0.36	*M* = 0.15 *SD* = 0.36	*M* = 0.32 *SD* = 0.47	*M* = 0.32 *SD* = 0.47
	m3.5	Es wird mit einer (grossen) Menge gemessen.	*M* = 0.46 *SD* = 0.50	*M* = 0.57 *SD* = 0.50	*M* = 0.59 *SD* = 0.49	*M* = 0.59 *SD* = 0.49
	m3.7	Der Wert der Menge wird auf die gesuchte Grösse zurückgerechnet.	*M* = 0.28 *SD* = 0.45	*M* = 0.28 *SD* = 0.45	*M* = 0.41 *SD* = 0.49	*M* = 0.41 *SD* = 0.49

	m3.8	Das Messen mit einer (noch grösseren) Menge wird als Lösungsvorschlag zur Steigerung der Messgenauigkeit angegeben.	*M* = 0.09 *SD* = 0.29	*M* = 0.09 *SD* = 0.29	*M* = 0.17 *SD* = 0.37	*M* = 0.17 *SD* = 0.37
QS 4 (Messinstrument)	m4.1	Das genauere Messinstrument (MI) wird erkannt. Die Begründung ist korrekt.	***M* = 0.52** *SD* = 0.50	***M* = 0.52** *SD* = 0.50	***M* = 0.86** *SD* = 0.35	***M* = 0.86** *SD* = 0.35
	m4.2	Es wird deutlich, dass für die Lösung mit dem genaueren MI gemessen wurde.	*M* = 0.85 *SD* = 0.36	*M* = 0.92 *SD* = 0.28	*M* = 0.94 *SD* = 0.25	*M* = 0.94 *SD* = 0.25
	m4.3	Das Messen mit einem (noch) genaueren MI wird als Lösungsvorschlag zur Steigerung der Messgenauigkeit angegeben.	*M* = 0.15 *SD* = 0.36	*M* = 0.15 *SD* = 0.36	*M* = 0.30 *SD* = 0.46	*M* = 0.30 *SD* = 0.46

Anmerkung: In Tabelle 11 sind die bepunkteten Indikatoren von QS 3 und QS 4 aufgeführt. Die Indikatoren wurden mit 'erfüllt' (1 Punkt) bzw. 'nicht erfüllt' (0 Punkte) bewertet. Fett hervorgehoben sind Mittelwerte, die sich beachtlich zwischen P, PV, PI und PVI unterscheiden.

Beim Betrachten der Mittelwerte der Indikatoren von QS 3 und QS 4 (vgl. Tab. 11) kann festgestellt werden, dass sich die Mittelwerte von P und PV vor allem bei den Indikatoren m3.1 und m4.1 von denjenigen von PI und PVI zu unterscheiden scheinen.

Um zu untersuchen, ob die Mittelwertsunterschiede zwischen den Ergebnissen der Kompetenzdiagnose hauptsächlich bei diesen zwei Indikatoren entstehen, wird eine weitere Analyse durchgeführt, in der diese beiden Indikatoren ausgeschlossen und dann die Mittelwerte und Mittelwertsunterschiede für die restlichen Indikatoren erneut betrachtet werden. In Tabelle 12 sind die Mittelwerte und Mittelwertsunterschiede der Ergebnisse der Kompetenzdiagnose vor dem Ausschluss (0 bis 15 Punkte möglich, vgl. auch Tab. 9) und nach dem Ausschluss der Indikatoren m3.1 und m4.1 (somit 0 bis 13 Punkte möglich) aufgeführt. Tabelle 12 zeigt, dass die Effekte der Mittelwertsunterschiede nach Ausschluss der Indikatoren m3.1 und m4.1 etwas geringer ausfallen. Dennoch kann festgestellt werden, dass sich die Mittelwerte von P und PI sowie P und PVI trotz Ausschluss der Indikatoren m3.1 und m4.1 bedeutsam unterscheiden ($|d| > 0.5$). Somit muss davon ausgegangen werden, dass auf Ebene der Stichprobe neben diesen

beiden Indikatoren noch weitere Indikatoren zu den Unterschieden zwischen den Ergebnissen der Kompetenzdiagnose führen. Aus diesem Grund werden Schülerprotokolle und ein Nachfragen lediglich bei den Indikatoren m3.1 und m4.1, zum Beispiel mit Hilfe eines Interviews, nicht für ein möglichst genaues Ergebnis der Kompetenzdiagnose ausreichen.

Tabelle 12: Mittelwerte und Mittelwertsunterschiede der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI aller bepunkteten Indikatoren des Problemtyps «Messen» (0 bis 15 Punkte möglich) und nach Ausschluss der Indikatoren m3.1 und m4.1 (0 bis 13 Punkte möglich); N = 108.

<table>
<tr><th></th><th>Mittelwerte (M) und Standardabweichungen (SD)</th><th>Prüfen der Mittelwertsunterschiede</th></tr>
<tr><td>Alle bepunkteten Indikatoren
(0 bis 15 Punkte möglich)</td><td>
<table>
<tr><th></th><th>M</th><th>SD</th></tr>
<tr><td>P</td><td>8.30</td><td>2.71</td></tr>
<tr><td>PV</td><td>8.63</td><td>2.48</td></tr>
<tr><td>PI</td><td>10.24</td><td>2.34</td></tr>
<tr><td>PVI</td><td>10.28</td><td>2.32</td></tr>
</table>
</td><td>
<table>
<tr><td>P vs. PV</td><td>t = -5.40</td><td>p ≤ .001</td><td>d = -0.13</td></tr>
<tr><td>P vs. PI</td><td>t = -11.92</td><td>p ≤ .001</td><td>d = -0.77</td></tr>
<tr><td>P vs. PVI</td><td>t = -12.14</td><td>p ≤ .001</td><td>d = -0.79</td></tr>
<tr><td>PV vs. PVI</td><td>t = -12.09</td><td>p ≤ .001</td><td>d = -0.69</td></tr>
<tr><td>PI vs. PVI</td><td>t = -1.42</td><td>p = .16 (n.s.)</td><td>-</td></tr>
</table>
</td></tr>
<tr><td>Nach Ausschluss der Indikatoren m3.1 und m4.1
(0 bis 13 Punkte möglich)</td><td>
<table>
<tr><th></th><th>M</th><th>SD</th></tr>
<tr><td>P</td><td>7.23</td><td>2.35</td></tr>
<tr><td>PV</td><td>7.56</td><td>2.12</td></tr>
<tr><td>PI</td><td>8.54</td><td>2.09</td></tr>
<tr><td>PVI</td><td>8.56</td><td>2.07</td></tr>
</table>
</td><td>
<table>
<tr><td>P vs. PV</td><td>t = -5.40</td><td>p ≤ .001</td><td>d = -0.15</td></tr>
<tr><td>P vs. PI</td><td>t = -9.44</td><td>p ≤ .001</td><td>d = -0.59</td></tr>
<tr><td>P vs. PVI</td><td>t = -9.60</td><td>p ≤ .001</td><td>d = -0.60</td></tr>
<tr><td>PV vs. PVI</td><td>t = -8.90</td><td>p ≤ .001</td><td>d = -0.48</td></tr>
<tr><td>PI vs. PVI</td><td>t = -1.42</td><td>p = .16 (n.s.)</td><td>-</td></tr>
</table>
</td></tr>
</table>

Anmerkungen: t-Test für abhängige Stichproben und Cohens d als Mass für die Effektstärken.

7.2.2 Vergleich der Ergebnisse der Kompetenzdiagnose auf individueller Ebene

Neben dem Vergleichen der Ergebnisse auf Ebene der Stichprobe (siehe Unterkapitel 7.2.1) ist auch von Interesse, inwiefern die Diagnose experimenteller Kompetenzen durch die verschiedenen Erhebungsmethoden (P, PV, PI und PVI) auf individueller Ebene, also auf Ebene der einzelnen Schülerinnen und Schülern, zu ähnlichen Ergebnissen gelangt. Hierfür werden die Zusammenhänge zuerst graphisch und dann mit Hilfe von Korrelationskoeffizienten betrachtet.

In Abbildung 11 werden die Zusammenhänge zwischen den Ergebnissen der Kompetenzdiagnose durch PVI (Benchmark) und den anderen Erhebungsmethoden (P, PV und PI) graphisch betrachtet, um zu untersuchen, inwiefern die Ergebnisse der Kompetenzdiagnose durch P, PV und PI bezüglich der Genauigkeit des Ergebnisses an den gesetzten Benchmark (PVI) herankommen und somit einen akzeptablen Ersatz für diesen bieten können (vgl. Kapitel 5). Insgesamt konnten bei den Aufgaben des Problemtyps «Messen» 0 bis 15 Punkte erreicht werden. Es kann festgestellt werden, dass die Kompetenzdiagnose durch PI und PVI auf Individualebene zu sehr ähnlichen Ergebnissen führt (vgl. Abb. 11, in Grau). Des Weiteren kann festgestellt werden, dass die experimentellen Kompetenzen durch PVI tendenziell höher diagnostiziert werden als durch P (vgl. Abb. 11, in Blau) und PV (vgl. Abb. 11, in Orange).

Abbildung 12 zeigt zudem den Zusammenhang zwischen den Ergebnissen der Kompetenzdiagnose durch P und PV (vgl. Abb. 12, in Orange) und P und PI (vgl. Abb. 12, in Grau), um der Frage nachzugehen, inwiefern *zusätzliche* Erhebungsmethoden zum Schülerprotokoll die Genauigkeit der Diagnostik erhöhen (vgl. Kapitel 5). Es kann festgestellt werden, dass die Ergebnisse der Kompetenzdiagnose durch PV leicht höher und durch PI einiges höher sind als die Ergebnisse der Kompetenzerfassung durch lediglich P.

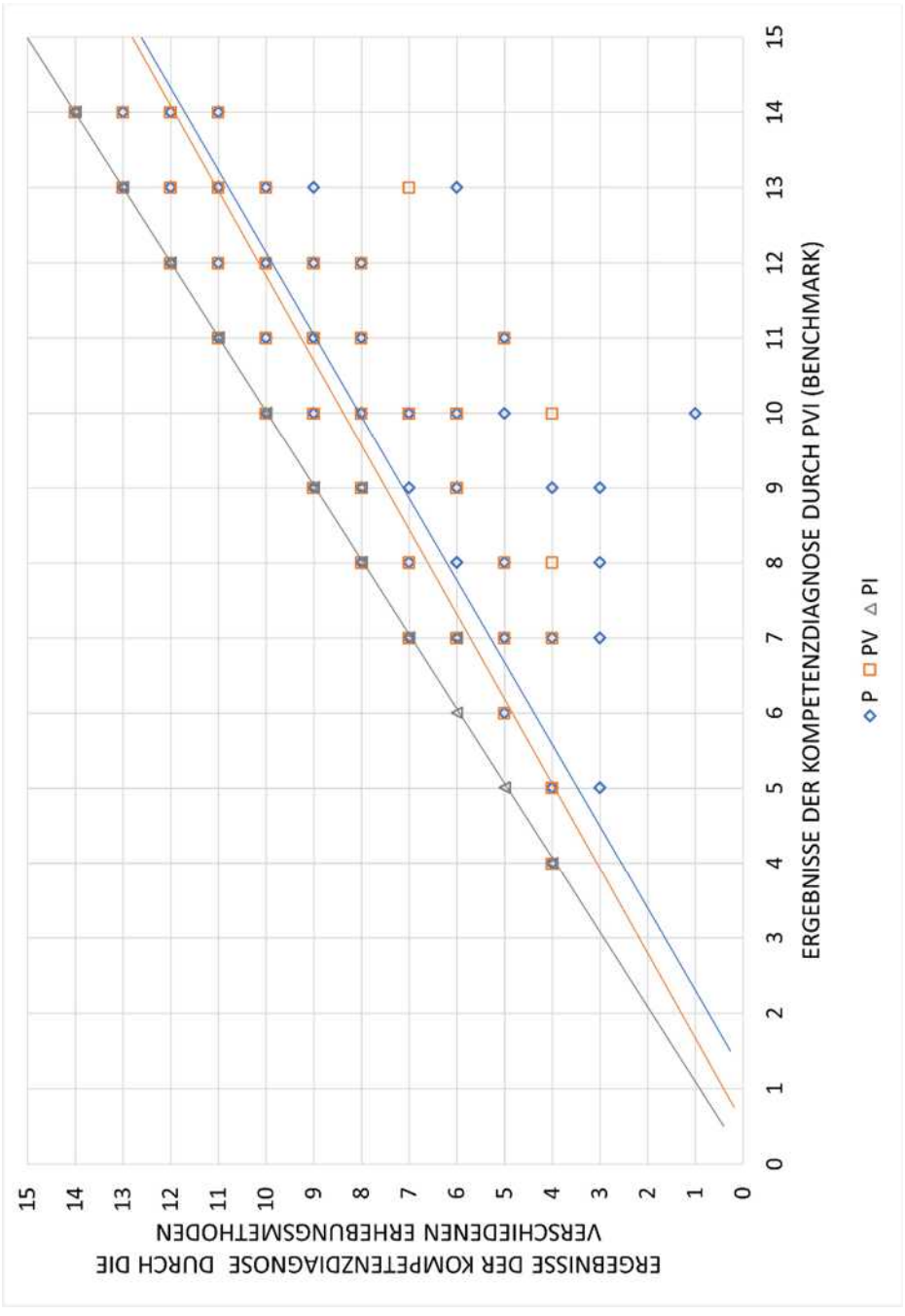


Abbildung 11: Graphische Betrachtung der Zusammenhänge zwischen den Ergebnissen der Kompetenzdiagnose durch PVI (Benchmark) und den anderen Erhebungsmethoden (P, PV und PI) bei den Aufgaben des Problemtyps «Messen» (0 bis 15 Punkte möglich, N = 108). Anmerkungen: Teilweise überlagern sich die Symbole (◊, □, Δ) oder können mehr als einen Schüler bzw. eine Schülerin repräsentieren (z. B. repräsentiert das Symbol □ bei Datenpunkt (5,4) drei Lernende). Die eingefügten linearen Trendlinien basieren auf allen Datenpunkten.

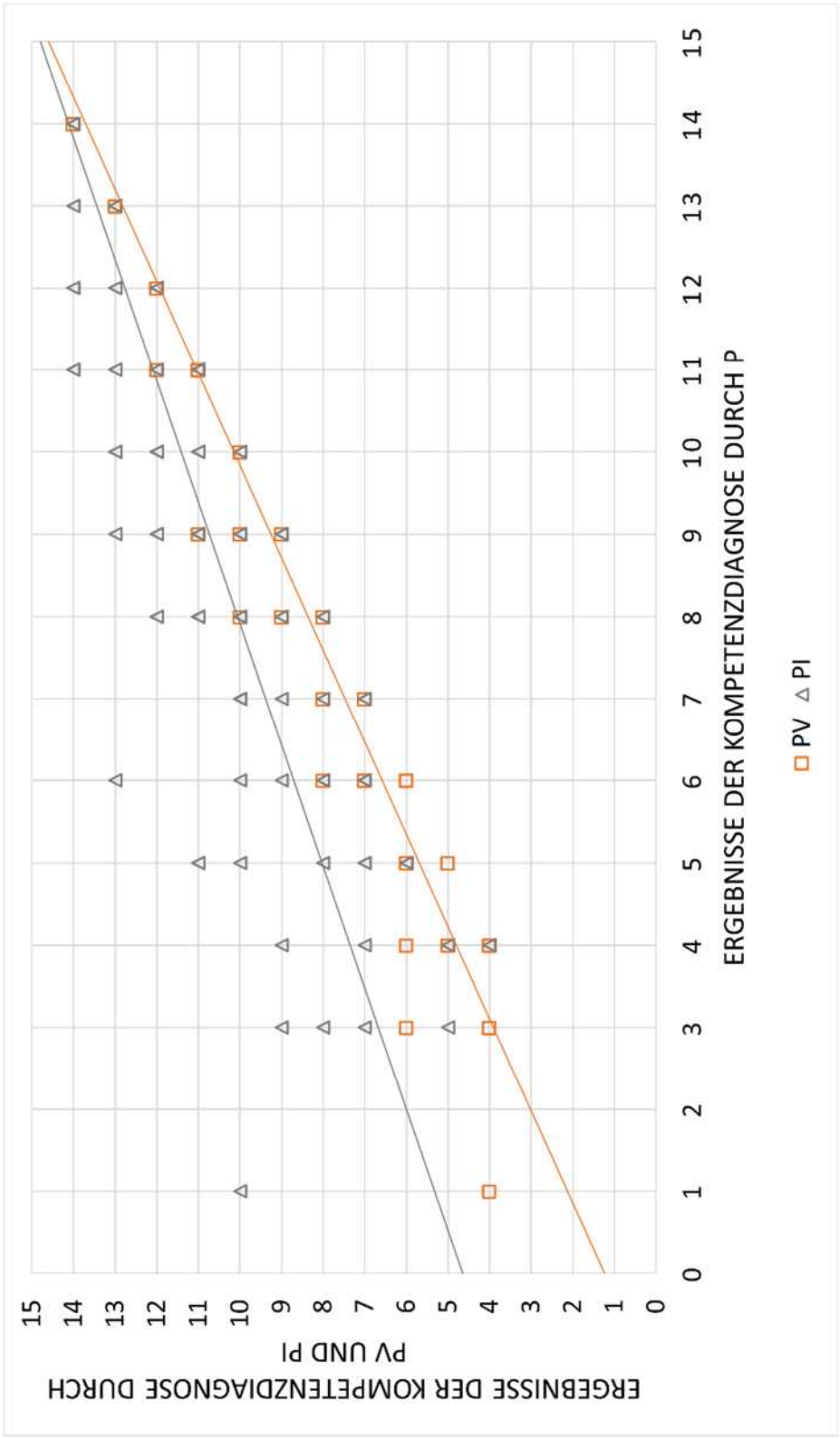


Abbildung 12: Graphische Betrachtung der Zusammenhänge zwischen den Ergebnissen der Kompetenzdiagnose durch P und PV sowie P und PI bei den Aufgaben des Problemtyps «Messen» (0 bis 15 Punkte möglich, N = 108). Anmerkungen: Teilweise überlagern sich die Symbole (□, Δ) oder repräsentieren mehr als einen Schüler bzw. eine Schülerin (z. B. repräsentiert das Symbol □ bei Datenpunkt (5,5) fünf Lernende). Die eingefügten linearen Trendlinien basieren auf allen Datenpunkten.

Die graphischen Betrachtungen der Zusammenhänge zwischen den Ergebnissen der Kompetenzdiagnose zeigen, dass PI auf individueller Ebene zu einem sehr ähnlichen Ergebnis der Diagnose gelangt wie der Benchmark (PVI, vgl. Abb. 11), und dass die experimentellen Kompetenzen durch zusätzliche Interviews (also PI bzw. PVI, vgl. Abb. 11 und 12) tendenziell höher diagnostiziert werden. In der Folge werden die Zusammenhänge mit Hilfe des Masses Kendall-Tau-b untersucht. Kendall-Tau-b ist ein Korrelationsmass für mindestens ordinalskalierte Daten, bei welchen Rangbindungen auftreten. Von Rangbindungen wird gesprochen, wenn in einer oder beiden Beobachtungsreihen gleiche Messwerte auftreten (Bortz & Lienert, 2008). Beim Auftreten von Rangbindungen ist der Korrelationskoeffizient Kendall-Tau-b dem Korrelationskoeffizienten Spearmans Rho vorzuziehen. In vorliegender Arbeit wird mit Hilfe Kendall-Tau-b geprüft, inwiefern auf Individualebene zwischen den Erhebungsmethoden (P, PV, PI und PVI) ein hinreichend hoher Zusammenhang besteht, sodass davon ausgegangen werden kann, dass die Erhebungsmethoden austauschbar sind. Da bei den Erhebungsmethoden eine Auseinandersetzung mit den gleichen Aufträgen stattgefunden hat (vgl. Schülerprotokolle, Unterkapitel 6.4.2) und die Diagnose experimenteller Kompetenzen mit sehr ähnlichen Kodiermanualen (vgl. Unterkapitel 7.1.1) vorgenommen wurde, werden im Rahmen der Auswertung relativ hohe Korrelationskoeffizienten erwartet. Dabei wurde im Rahmen vorliegender Studie eine Korrelation von $r > .7$ als ausreichend hoch definiert, um von einer Austauschbarkeit zu sprechen, analog zur Studie von Schreiber (2012) zur Substituierbarkeit von Testarten. In Tabelle 13 werden die Korrelationen für alle bepunkteten Indikatoren des Problemtyps «Messen» sowie nach Ausschluss der Indikatoren m3.1 und m4.1 betrachtet, da sich die Mittelwerte der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI bei diesen beiden Indikatoren in der Einzelbetrachtung auffällig unterschieden (vgl. Tab. 11).

Tabelle 13: Korrelationen zwischen P, PV, PI und PVI aller bepunkteten Indikatoren des Problemtyps «Messen» und nach Ausschluss der Indikatoren m3.1 und m4.1; N = 108.

	Alle bepunkteten Indikatoren		**Nach Ausschluss der Indikatoren m3.1 und m4.1**	
Korrelationen Kendall-Tau-b	P und PV	$r = .94$ $p \leq .001$	P und PV	$r = .92$ $p \leq .001$
	P und PI	$r = .67$ $p \leq .001$	P und PI	$r = .70$ $p \leq .001$
	P und PVI	$r = .67$ $p \leq .001$	P und PVI	$r = .70$ $p \leq .001$
	PV und PVI	$r = .71$ $p \leq .001$	PV und PVI	$r = .74$ $p \leq .001$
	PI und PVI	$r = .99$ $p \leq .001$	PI und PVI	$r = .99$ $p \leq .001$

Tabelle 13 zeigt, dass die Korrelationen zwischen den Ergebnissen der Kompetenzdiagnose durch P und PV sowie PI und PVI sehr hoch ausfallen und diese Methoden somit auf Individualebene zu einem sehr ähnlichen Ergebnis der Kompetenzdiagnose gelangen. Die Korrelationen zwischen den Ergebnissen der Kompetenzdiagnose durch P und PI, P und PVI sowie PV und PVI erfüllen, wenn überhaupt, dann nur sehr knapp die festgelegten Bedingungen für eine hinreichend hohe Korrelation ($r > .7$), weshalb im Rahmen vorliegender Arbeit, auch unter Einbezug der Abbildungen 11 und 12, davon ausgegangen wird, dass die Erhebungsmethoden auf Individualebene zu unterschiedlichen Ergebnissen der Kompetenzdiagnose führen. Somit scheinen zusätzliche Interviews auch auf Individualebene die Genauigkeit der Diagnose zu erhöhen.

Insgesamt deuten die Ergebnisse von Teilstudie I darauf hin, dass zusätzliche Interviews die Genauigkeit der Ergebnisse der Kompetenzdiagnose erhöhen. Infolgedessen stellt sich die Frage, ob die genauere Diagnose möglicherweise mit bestimmten Variablen (z. B. Aufgabenkontext, Messzeitpunkt oder spezifische Schülermerkmale) erklärt werden kann. Um dies ansatzweise zu begründen, wurde qualitativ untersucht, ob grössere Differenzen im Ergebnis der Kompetenzdiagnose durch zusätzliche Interviews im Vergleich zu ohne Interviews (P vs. PI) augenscheinlich mit anderen Variablen zusammenhängen. Als grössere Differenzen wurden dabei Differenzen von mindestens drei Punkten betrachtet, da festgestellt werden konnte, dass im Mittel durch zusätzliche Interviews die experimentellen Kompetenzen der Lernenden um circa zwei Punkte höher diagnostiziert werden (vgl. Tab. 9). Insgesamt konnten 34 Fälle (von $N = 108$) festgestellt werden, bei welchen die Differenz im Ergebnis der Kompetenzdiagnose durch zusätzliche Interviews im Vergleich zu ohne Interviews mindestens drei Punkte betrug. Da es plausibel erscheint, dass diese Differenzen womöglich mit dem Aufgabenkontext, dem Messzeitpunkt oder gewissen Schülermerkmalen (z. B. sprachliche / kognitive Fähigkeiten oder Motivation) zusammenhängen könnten, wurden diese möglichen Korrelationen qualitativ auf der Ebene einzelner Schülerinnen und Schülern betrachtet. Dabei konnte festgestellt werden, dass die 34 Fälle über alle Aufgaben des Problemtyps «Messen» hinweg verteilt waren: Somit scheinen grössere Differenzen im Ergebnis der Kompetenzdiagnose durch zusätzliche Interviews nicht vom Kontext der Aufgabe abhängig zu sein. Bezüglich des Messzeitpunkts konnte festgestellt werden, dass die 34 Fälle vermehrt (bei 22 von 34 Fällen) bei Besuch 1 und 2[32] vorkamen, womit der Messzeitpunkt einen Einfluss auf die genauere Diagnose experimenteller

[32] Vgl. Ablauf der Datenerhebung, Unterkapitel 6.3.

Kompetenzen anhand zusätzlicher Interviews zu haben scheint. Zudem sind die 34 Fälle über die Mehrheit der Schülerinnen und Schüler (21 von 27 Lernenden) verteilt. Somit scheinen spezifische Schülermerkmale (z. B. sprachliche / kognitive Fähigkeiten oder Motivation) nicht einen entscheidenden Einfluss auf eine Differenz von *mindestens drei Punkten* zu haben. Wird hingegen betrachtet, bei welchen Schülerinnen und Schülern die Differenz durch zusätzliche Interviews besonders stark ausgeprägt ist (Differenzen von P und PI ≥ 5 Punkte; insgesamt bei 8 Fällen und 6 von 27 Lernenden), konnte festgestellt werden, dass diese Schülerinnen und Schüler vermehrt schwächere Leistungen in den sprachlichen und kognitiven Begleittests sowie tiefere Werte beim Fragebogen zur Motivation zeigten (ein/e SchülerIn: tiefe Leistungen in den sprachlichen Begleittests und tiefe Werte beim Motivationsfragebogen; ein/e SchülerIn: tiefe Leistungen im kognitiven Begleittest und tiefe Werte beim Motivationsfragebogen; zwei SchülerInnen: tiefe Leistungen in den sprachlichen Begleittests und im kognitiven Begleittest)[33]. Somit scheinen sprachliche und kognitive Prädispositionen sowie motivationale Faktoren gemäss der qualitativen Betrachtung einen Einfluss auf *eine besonders stark ausgeprägte Differenz* (Differenz P und PI ≥ 5 Punkte) im Ergebnis der Kompetenzdiagnose zu haben. Im Interview scheinen bei diesen Lernenden die sprachlichen und kognitiven Schwierigkeiten (z. B. Probleme beim Schreiben eines Protokolls, kognitive Überlastung) sowie motivationale Faktoren (z. B. keine Lust zum Schreiben eines Protokolls) weniger Auswirkung zu haben, wodurch anhand eines zusätzlichen Interviews eindeutig höhere Ergebnisse bei der Kompetenzdiagnose erzielt werden.

7.3 Fazit und Diskussion

Teilstudie I untersucht, inwiefern die Ergebnisse der Kompetenzdiagnose durch verschiedene Erhebungsmethoden (P, PV und PI) bezüglich der Genauigkeit des Ergebnisses der Diagnose an den gesetzten Benchmark (PVI) herankommen und inwiefern *zusätzliche* Erhebungsmethoden zum Schülerprotokoll (P vs. PV, P vs. PI) die Genauigkeit der Diagnostik erhöhen. In Unterkapitel 7.2 wurden die Ergebnisse zum Vergleich der Ergebnisse der Kompetenzdiagnose durch P, PV, PI und PVI auf Ebene der Stichprobe und auf Individualebene aufgeführt. In diesem Unterkapitel werden die Ergebnisse zusammengefasst und diskutiert, wobei auch Bezug zu den hergeleiteten Hypothesen (vgl. Kapitel 5) genommen wird.

[33] Begleittests (vgl. auch Unterkapitel 6.3): Sprachliche Begleittests, KFT Teilbereich V und SLS; kognitiver Begleittest, KFT Teilbereiche Q und N; Fragebogen Motivation und situationales Interesse.

Die Hypothesen H1 und H2 gehen davon aus, dass zusätzliche Videoaufnahmen von den Schülerinnen und Schülern während des Experimentierens für ein möglichst genaues Ergebnis der Kompetenzdiagnose nicht zwingend benötigt werden und somit die Ergebnisse der Kompetenzdiagnose durch P und PV (vgl. H1) sowie durch PI und PVI (vgl. H2) auf Ebene der Stichprobe und Individualebene sehr ähnlich und somit die Methoden grundsätzlich austauschbar sind. Die Hypothesen H1 und H2 können bestätigt werden. Auf Ebene der Stichprobe konnte festgestellt werden, dass sich die Mittelwerte der Ergebnisse der Kompetenzdiagnose durch P und PV zwar signifikant unterscheiden, der Effekt jedoch nur als sehr gering einzustufen ist ($p \leq .001$; $d = -0.13$). Die Mittelwertsunterschiede zwischen den Ergebnissen der Kompetenzdiagnose durch PI und PVI fallen zudem nicht signifikant aus ($p = .16$), was für die Austauschbarkeit der Methoden auf Ebene der Stichprobe spricht. Zudem konnten auf Individualebene hohe Korrelationen zwischen P und PV ($r = .94$; $p \leq .001$) sowie PI und PVI ($r = .99$; $p \leq .001$) beobachtet werden, womit die Erhebungsmethoden zu einem ähnlichen Ergebnis der Kompetenzdiagnose gelangen und somit grundsätzlich austauschbar sind. Die Bestätigung der Hypothesen H1 und H2 auf Ebene der Stichprobe und auf Individualebene zeigen, dass zusätzliche Videoaufnahmen von Schülerinnen und Schülern des 8. Schuljahres während des Experimentierens für ein möglichst genaues Ergebnis der Kompetenzdiagnose bei den Aufgaben des Problemtyps «Messen» nicht zwingend notwendig scheinen und somit PI bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose an den gesetzten Benchmark (PVI) heranzukommen scheint. Dies kann für Lehrpersonen des naturwissenschaftlichen Unterrichts sowie für zukünftige Studien eine Entlastung bedeuten: Videoaufnahmen scheinen gemäss dieser Studie keinen entscheidenden Vorteil bezüglich der Genauigkeit des Ergebnisses der Diagnose zu bringen, insbesondere dann nicht, wenn zusätzliche Interviews durchgeführt wurden. Bezüglich des naturwissenschaftlichen Unterrichts könnte man diese Erkenntnis auch auf die direkte Beobachtung von Schülerinnen und Schülern während des Experimentierens erweitern. Dass Videoaufnahmen respektive das Beobachten von Schülerinnen und Schülern während des Experimentierens keinen grossen Vorteil bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose bringen und sich durch Schülerprotokolle substituieren lassen, zeigen auch andere Studien (vgl. z. B. Baxter et al., 1992; Emden & Sumfleth, 2012; Shavelson et al., 1991, 1993). Somit leisten die Ergebnisse von Teilstudie I auch einen zentralen Beitrag zur Bestätigung bestehender Befunde. Dennoch bedeuten die Ergebnisse nicht, dass nun vollständig auf das Beobachten von Schülerinnen und Schülern während des Experimentierens verzichtet werden kann, denn (1) wurden im Rahmen vorliegender Studie leistungsstärkere Schülerinnen und Schüler

untersucht und bei leistungsschwächeren Lernenden kann das Beobachten durchaus sinnvoll sein, weil ihnen womöglich das Führen eines Protokolls nicht gelingt (vgl. z. B. auch Emden & Sumfleth, 2012; Gott & Duggan, 2002) und (2) kann die Beobachtung von Schülerinnen und Schülern während des Experimentierens an manchen Stellen trotz des erhöhten Aufwands sinnvoll sein, weil dadurch mehr über handlungsbezogene experimentelle Kompetenzen, wie zum Beispiel die Handhabung von Messinstrumenten, in Erfahrung gebracht werden kann.

Die Hypothesen H3 bis H5 gehen davon aus, dass zusätzliche Interviews auf Ebene der Stichprobe und Individualebene zu einem genaueren Ergebnis der Kompetenzdiagnose führen. Somit sollten sich die Ergebnisse der Kompetenzdiagnose durch P und PI (vgl. H3), P und PVI (vgl. H4) sowie PV und PVI (vgl. H5) bedeutsam bezüglich der Genauigkeit des Ergebnisses der Diagnose unterscheiden. Gemäss den Ergebnissen von Teilstudie I können die Hypothesen H3, H4 und H5 bestätigt werden. Auf Ebene der Stichprobe können bei den Ergebnissen der Kompetenzdiagnose durch P und PI ($p \leq .001$; $d = -0.77$), P und PVI ($p \leq .001$; $d = -0.79$) sowie PV und PVI ($p \leq .001$; $d = -0.69$) signifikante und bedeutsame ($|d| > 0.5$) Mittelwertsunterschiede beobachtet werden. Auf Individualebene konnte zudem festgestellt werden, dass die Ergebnisse der Kompetenzdiagnose durch P und PI ($r = .67$; $p \leq .001$), P und PVI ($r = .67$; $p \leq .001$) sowie PV und PVI ($r = .71$; $p \leq .001$) nicht (oder nur sehr knapp) hinreichend hoch ($r > .7$) miteinander korrelieren. Somit scheinen zusätzliche Interviews auf Ebene der Stichprobe und auf Individualebene zu einem genaueren Ergebnis der Kompetenzdiagnose zu führen und die Ergebnisse der Kompetenzdiagnose durch P und PI, P und PVI sowie PV und PVI nicht austauschbar zu sein. Dieser Befund deckt sich mit den Erwartungen, die aus der Literatur hergeleitet werden können. Interviews ermöglichen einen Zugang zu kognitiven Prozessen (vgl. Funke & Spering, 2006) und somit kann mehr über die Gedanken und Überlegungen von Schülerinnen und Schülern während des Experimentierens in Erfahrung gebracht werden. Zudem weisen Studien darauf hin, dass zwischen schriftlichen Materialien und den Gedanken von Schülerinnen und Schülern nicht stets ein systematischer Zusammenhang besteht (vgl. z. B. Vorholzer et al., 2020) und Schülerprotokolle nur einen eingeschränkten Einblick über die tatsächlichen experimentellen Kompetenzen zulassen (vgl. z. B. Abrahams et al. 2013).
Die genauere Diagnose experimenteller Kompetenzen durch zusätzliche Interviews könnte teilweise durch die unterschiedlichen Charakteristika der Erhebungsmethoden begründet werden. Während im Schülerprotokoll öfters Aufträge unbeantwortet bleiben und somit in diesen Bereichen bei der Diagnose keine Punkte erzielt werden, kommen in den Interviews unbeantwortete Fragen kaum

vor. Dies liegt daran, dass die Lernenden in einem Interview den Aufträgen respektive Fragen weniger ‘ausweichen’ können. Zudem erhalten die Schülerinnen und Schüler in einem Interview durch das Nachfragen zusätzliche Stimuli. Diese Stimuli könnten die Lernenden zu Gedanken anregen, die sie sich zuvor noch nicht gemacht haben. Diese Charakteristika der Erhebungsmethoden könnten unter anderem dazu führen, dass durch zusätzliche Interviews die experimentellen Kompetenzen genauer (bzw. höher) diagnostiziert werden. Zudem konnte im Rahmen der qualitativen Betrachtung festgestellt werden, dass der Messzeitpunkt einen Einfluss auf die genauere Diagnose experimenteller Kompetenzen anhand zusätzlicher Interviews zu haben scheint, denn grössere Differenzen im Ergebnis der Kompetenzdiagnose mit und ohne zusätzliche Interviews konnten vor allem bei den ersten Schulbesuchen festgestellt werden. Dies könnte auf eine Gewöhnung an die Protokollmethode hindeuten (vgl. auch Emden und Sumfleth, 2012), wodurch bei den späteren Schulbesuchen anscheinend die Schülerinnen und Schüler die Schülerprotokolle genauer führen und somit die Ergebnisse der Kompetenzdiagnose mit und ohne zusätzliche Interviews besser übereinstimmen. Des Weiteren hat die qualitative Betrachtung gezeigt, dass vor allem bei leistungsschwächeren Schülerinnen und Schülern und / oder Lernenden mit geringer Motivation zusätzliche Interviews einen entscheidenden Vorteil auf die Genauigkeit der Diagnostik zu haben scheinen, denn diese Jugendlichen erzielen durch zusätzliche Interviews eindeutig höhere Ergebnisse bei der Kompetenzdiagnose. Im Rahmen vorliegender Arbeit wurden die Zusammenhänge jedoch nur qualitativ betrachtet und somit kann nicht beurteilt werden, inwiefern die genauere Diagnose durch zusätzliche Interviews womöglich hauptsächlich von den kognitiven respektive sprachlichen Fähigkeiten oder der Motivation abhängt. Hier bedarf es weiterer Forschung, indem diese möglichen Zusammenhänge systematisch untersucht werden (vgl. Kapitel 9). Eine weitere mögliche Ursache für die genauere (bzw. höhere) Diagnose experimenteller Kompetenzen durch zusätzliche Interviews könnte im Phänomen der sozialen Erwünschtheit liegen. Dies bedeutet, dass sich die Antworten der Lernenden nicht an dem subjektiv ‘wahren’ Wert orientieren, sondern an den im Interview wahrgenommenen Erwartungen (vgl. Stocké, 2019). Das Phänomen der sozialen Erwünschtheit kann bei Interviews jedoch kaum umgangen werden. Trotz dieser Einschränkung ist aber aufgrund der Ergebnisse von Teilstudie I davon auszugehen, dass zusätzliche Interviews zu einem genaueren Ergebnis der Kompetenzdiagnose führen. Dieser Befund ist von zentraler Bedeutung für den naturwissenschaftlichen Unterricht. Er zeigt, dass es trotz des organisatorischen und zeitlichen Zusatzaufwands sinnvoll und notwendig ist, die Lernenden manchmal zu ihren experimentellen Handlungen zu befragen. So kann mehr über experimentelle Handlungen, die womöglich

nicht ausreichend in den Schülerprotokollen festgehalten wurden, und über kognitive Prozesse seitens der Lernenden in Erfahrung gebracht werden. Die Ergebnisse von Teilstudie I sollen den Einsatz von lediglich Schülerprotokollen aber nicht ‘verbieten’. In manchen Situationen sind Schülerprotokolle aus testökonomischen Gründen die beste Wahl. Man muss sich jedoch bewusst sein, dass es eine Einführung in die Protokollmethode bedarf (Gewöhnungseffekt, vgl. auch Emden und Sumfleth, 2012) und dass zusätzliche Interviews, vor allem auch bei leistungsschwächeren Schülerinnen und Schülern oder Lernenden mit geringer Motivation, zu einem genaueren Ergebnis der Kompetenzdiagnose führen könnten.

8. Teilstudie II: Kognitive Validierung am Beispiel der Aufgaben des Problemtyps «Messen»

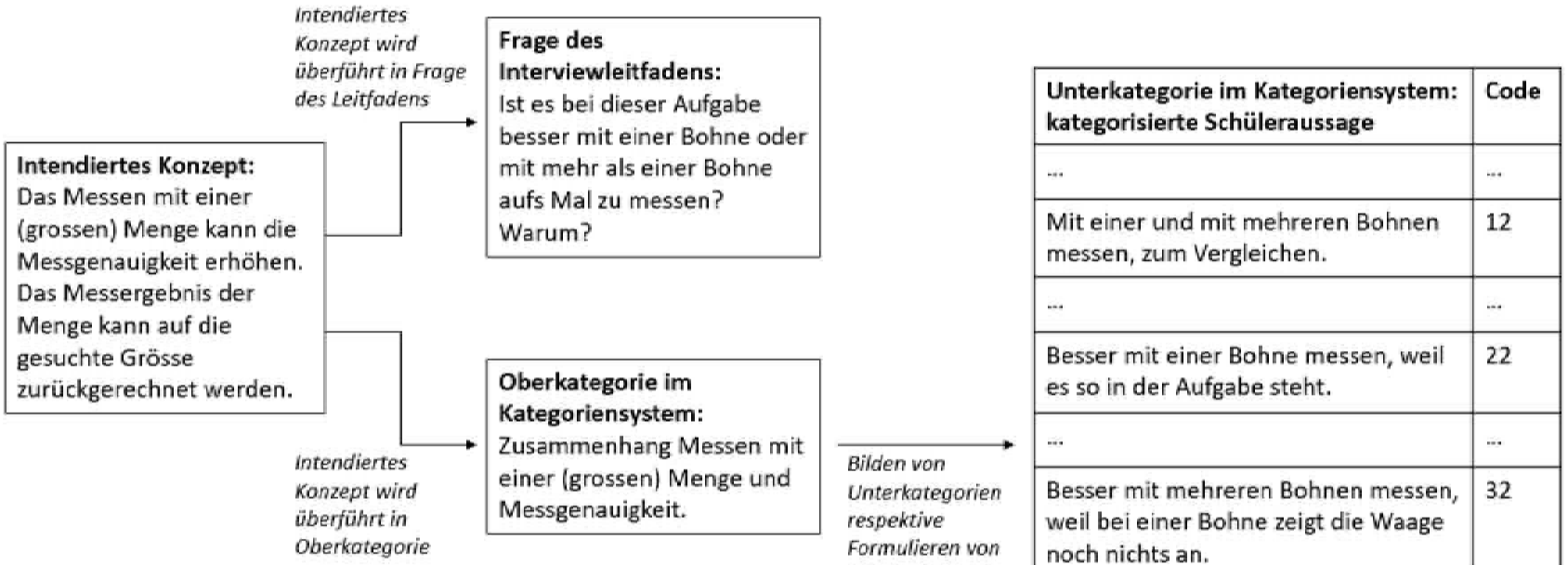
Im Rahmen von Teilstudie II wird Forschungsfrage 2 (FF2, vgl. Kapitel 5) nachgegangen, indem untersucht wird, inwiefern die Aufgaben des Problemtyps «Messen» kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen. Somit leistet Teilstudie II einen zentralen Beitrag zu einer umfassenden Validitätsbeurteilung im Rahmen der Gesamtvalidierungsstudie des Projekts (vgl. Unterkapitel 6.1).
Um FF2 zu beantworten, wurden zum Teil die gleichen Daten wie in Teilstudie I genutzt: Vor allem die Daten der Interviews mit den Schülerinnen und Schülern und teilweise auch die Daten der Schülerprotokolle. Der Aufbau der Schülerprotokolle und Interviews wurde bereits in Unterkapitel 6.4 erläutert, somit liegt in diesem Kapitel der Fokus auf der Auswertung und Nutzung der Daten im Rahmen von Teilstudie II. Zudem wurde im Rahmen von Teilstudie II ein Expertenrating[34] durchgeführt, um unter anderem zu untersuchen, ob die Schülerinnen und Schüler beim Lösen der Aufgaben mehrheitlich über die intendierten Konzepte nachdenken. Für die Entwicklung des Expertenratings wurden die kategorisierten Schüleraussagen[35] aus den Interviews verwendet. Dafür mussten die Interviews in einem ersten Schritt mit Hilfe eines Kategoriensystems ausgewertet werden. In den folgenden Ausführungen wird folglich zuerst die Auswertung der Interviews mit Hilfe eines Kategoriensystems beschrieben und daraufhin wird auf das Expertenrating eingegangen. Danach werden die Ergebnisse im Rahmen von Teilstudie II erläutert und diskutiert.

8.1 Auswertung der Interviews mit Hilfe eines Kategoriensystems

Um FF2 zu untersuchen wurden die Interviews mit Hilfe eines Kategoriensystems ausgewertet. Die Auswertung mit Hilfe eines Kategoriensystems ermöglicht zu erfassen, worüber die Lernenden beim Lösen der Aufgaben des Problemtyps «Messen» hauptsächlich nachdenken. Dadurch kann untersucht werden, inwiefern die Gedanken und Überlegungen der Lernenden den intendierten Konzepten entsprechen. Falls diese mehrheitlich den intendierten Konzepten entsprechen, dann kann dies als Hinweis für kognitive Validität gedeutet werden: Die

[34] Eigentlich ist es ein Expertinnen- und Expertenrating, das für die bessere Lesbarkeit in vorliegender Arbeit als Expertenrating bezeichnet wird.
[35] Eigentlich sind es Schülerinnen- und Schüleraussagen, die für die bessere Lesbarkeit in vorliegender Arbeit als Schüleraussagen bezeichnet werden.

intendierten Konzepte werden somit anscheinend mehrheitlich seitens der Lernenden aktiviert.
Für jede Aufgabe des Problemtyps «Messen» wurde ein Kategoriensystem entwickelt, wobei die Kategoriensysteme der Aufgaben jeweils analog aufgebaut sind, das heisst die gleichen Oberkategorien umfassen. Die Oberkategorien wurden anhand der intendierten Konzepte der Aufgaben des Problemtyps «Messen» (vgl. Unterkapitel 6.4.1) entwickelt. Ausgehend von den intendierten Konzepten wurden in einem ersten Schritt entsprechende Fragen im Interviewleitfaden (vgl. Unterkapitel 6.4.4 und Anhang, Teil A) und daraufhin für die Auswertung der Interviews entsprechende Oberkategorien im Kategoriensystem entwickelt (vgl. Abb. 13). Die Bildung der Oberkategorien erfolgte somit deduktiv. Die Oberkategorien wurden im Zuge der Auswertung mit induktiv anhand des Datenmaterials entwickelten Unterkategorien, die den kategorisierten Schüleraussagen entsprechen, ergänzt. Die Unterkategorien wurden genutzt, um zu erfassen, worüber die Schülerinnen und Schüler beim Lösen der Aufgaben hauptsächlich nachdenken. Für die Auswertung der Interviews wurden daraufhin den Unterkategorien entsprechende Codes zugewiesen.



Unterkategorie im Kategoriensystem: kategorisierte Schüleraussage	Code
...	...
Mit einer und mit mehreren Bohnen messen, zum Vergleichen.	12
...	...
Besser mit einer Bohne messen, weil es so in der Aufgabe steht.	22
...	...
Besser mit mehreren Bohnen messen, weil bei einer Bohne zeigt die Waage noch nichts an.	32
Besser mit mehreren Bohnen messen, weil mit mehreren Bohnen wird es genauer.	33
Bohnen sind unterschiedlich gross / haben unterschiedliche Massen. Darum besser mit mehreren Bohnen messen.	34

Abbildung 13: Überführung der intendierten Konzepte in entsprechende Fragen im Interviewleitfaden und Oberkategorien im Kategoriensystem. Formulieren von Unterkategorien bzw. kategorisierten Schüleraussagen. Hier am Beispiel der Bohnenaufgabe im Bereich der Strategie Mengenvergrösserung.

Die Kodierung der Interviews mit Hilfe des Kategoriensystems fand von Juni bis August 2018 statt. Die Interviews wurden aus ökonomischen Gründen nicht transkribiert, sondern es wurden direkt die Videoaufzeichnungen der Interviews analysiert. Alle Interviews (N = 108) wurden zuerst von der Autorin vorliegender Arbeit analysiert. Dabei wurden induktiv die Unterkategorien gebildet. Für die Bildung der Unterkategorien wurden alle konkreten Äusserungen der Schülerinnen und Schüler, bei denen Vorgehensweisen oder Überlegungen im Bereich der intendierten Konzepte geäussert wurden, berücksichtigt. Nach der Entwicklung des Kategoriensystems wurde dieses genutzt und ein Teil der Daten doppelt kodiert, um eine objektive Kodierung zu gewährleisten. Vor Beginn der Doppelkodierung wurde eine Kodier-Schulung durchgeführt, bei der das Kategoriensystem erläutert und die Kodierung anhand von Beispielen erklärt wurde. 30 % der Interviews wurden doppelt kodiert, wobei eine zufriedenstellende Beurteilerübereinstimmung erreicht werden konnte (Cohens $k \geq .70$; prozentuale Übereinstimmung (pÜ) ≥ 78 %).

Bei der Kodierung der Interviews wurde die gesamte Antwort zu einer Frage des Interviewleitfadens betrachtet (Kodiereinheit) und dann die Antwort des Schülers respektive der Schülerin einer kategorisierten Schüleraussage (Unterkategorie) zugewiesen. Jeder Antwort zu einer Interviewfrage wurde jeweils nur eine Unterkategorie zugewiesen, und zwar diejenige, um die es gemäss der Einschätzung der Kodiererin bei der Antwort hauptsächlich ging. In Abbildung 14 wird die Zuordnung der Antwort eines Schülers zu einer Unterkategorie am Beispiel der Ahornaufgabe illustriert. Bei der Antwort des Schülers geht es einerseits um Ungenauigkeiten, die entstehen können (vgl. Abb. 14, *«…es kann sein, dass der Samen komisch fällt…»*) und zudem erwähnt der Schüler, dass die Zeit, die man für die Aufgabe zur Verfügung hat, auch eine Rolle spielt (vgl. Abb. 14, *«…wenn es die Zeit zulässt, 10mal oder sonst mindestens 5mal messen, aber sicher mehr als nur einmal.»*). Da es bei der Antwort des Schülers und somit bei der Begründung für das Durchführen von Messwiederholungen gemäss Einschätzungen der Kodiererinnen aber hauptsächlich um das Ausgleichen von Ungenauigkeiten geht, wurde dieser Schülerantwort der Code 28 zugewiesen.

Intendiertes Konzept: Durch das Durchführen von Messwiederholungen und das Bilden eines Mittelwerts kann die Messgenauigkeit erhöht werden.

Interviewfrage: Ist es bei dieser Aufgabe besser einmal oder mehrmals zu messen? Warum?

Antwort des Schülers: «Mehrmals messen ist besser, weil es kann sein, dass der Samen komisch fällt und nicht immer ganz genau gleich hinunterfällt. Darum würde ich, wenn es die Zeit zulässt, 10mal oder sonst mindestens 5mal messen, aber sicher mehr als nur einmal.»

Oberkategorie: Zusammenhang zwischen dem Durchführen von Messwiederholungen und der Messgenauigkeit.

Unterkategorien / kategorisierte Schüleraussagen

Einmal messen genügt, weil ...	Keine weitere Begründung / unvollständige Begründung / unklare Begründung.	Code 11
	... die Aufgabe nicht verlangt, dass man mehrmals misst.	Code 12
	... es wird sowieso in etwa das gleiche Ergebnis herauskommen.	Code 13
	... man die Grösse sowieso nicht genau messen kann. Ein ungefährer Wert genügt.	Code 14
	... wenn das Experiment genau durchgeführt wurde, dann genügt einmal messen.	Code 15
Mehrmals messen ist besser, weil ...	Keine weitere Begründung / unvollständige Begründung / unklare Begründung.	Code 21
	... man hat bei dieser Aufgabe genügend Zeit.	Code 22
	... andere, oft unlogische, Begründung (z. B. je mehr man misst, desto schneller fliegt der Samen; etc.).	Code 23
	... die zu messende Grösse schwierig zu messen ist.	Code 24
	... man sich so sicher sein kann / ein Ergebnis bestätigen kann.	Code 25
	... man so sehen kann in welchem Bereich der Wert liegt.	Code 26
	... man so Ausreisser erkennen und gegebenenfalls ausschliessen kann.	Code 27
	... sich so Ungenauigkeiten ausgleichen (z. B. Samen fallen nicht immer gleich).	Code 28
	... es genauer wird. Ohne weitere Begründung oder nicht verständliche Begründung, warum das so ist.	Code 29
	... man so ein genaueres Ergebnis kriegt und einen Mittelwert berechnen kann.	Code 30

Abbildung 14: Beispiel für eine Zuordnung einer Antwort eines Schülers zu einer Unterkategorie bzw. kategorisierten Schüleraussage am Beispiel der Ahornaufgabe im Bereich Messwiederholung.

Ein Beispiel eines Kategoriensystem ist im Anhang zu finden (vgl. Anhang, Teil B). Neben den Kategorien zu den intendierten Konzepten sind im Kategoriensystem auch Kategorien zu den allgemeineren Fragen des Interviews zu finden (vgl. Unterkapitel 6.4.4). Die Kodierung dieser Kategorien wurden im Zuge des Untersuchens von FF1 und FF2 nicht genutzt, kann aber der weiterführenden Forschung dienen (vgl. Kapitel 9).

8.2 Erhebung Expertenrating

Zum Lösen der Aufgaben des Problemtyps «Messen» müssen die Schülerinnen und Schüler gewisse Konzepte (z. B. das Durchführen von Messwiederholungen und die Bildung eines Mittelwerts kann die Messgenauigkeit erhöhen) zumindest intuitiv verstanden haben. Im Rahmen der kognitiven Validierung wird unter anderem untersucht, inwiefern diese Konzepte seitens der Schülerinnen und Schüler aktiviert werden und somit die Aufgaben kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich dieser Konzepte zulassen. Hinweise für kognitive Validität können dabei aus unterschiedlichen Bereichen stammen (vgl. Unterkapitel 4.2): (I) die intendierten Konzepte werden seitens der Schülerinnen und Schüler aktiviert und zum Lösen der Aufgaben verwendet, (II) qualitativ hochwertigere Denkprozesse gehen mit einer besseren Lösung der Aufgabe einher und (III) die Expertinnen und Experten schätzen die intendierten Konzepte als naheliegend ein und erkennen keine Aspekte, welche die kognitive Validität beeinträchtigen könnten. Ziel des Expertenratings ist es somit einerseits zu prüfen, inwiefern die intendierten Konzepte bei den Aufgaben überhaupt naheliegend sind und dadurch bei den Schülerinnen und Schülern aktiviert werden können (Hinweise aus dem Bereich (III)). Sollen beispielsweise die Lernenden über den Zusammenhang zwischen Messwiederholungen und der Messgenauigkeit nachdenken, dann sollte das Durchführen von Messwiederholungen bei der Aufgabe auch naheliegend sein. Andererseits wird durch die Einschätzungen der Expertinnen und Experten im Rating geprüft, inwiefern die Gedanken und Überlegungen der Schülerinnen und Schüler den intendierten Konzepten entsprechen. Hierzu wurden die kategorisierten Schüleraussagen der Auswertung der Interviews genutzt (vgl. Unterkapitel 8.1) und die Expertinnen und Experten schätzten diese im Hinblick darauf ein, inwiefern diese den intendierten Konzepten entsprechen. Falls so gezeigt werden kann, dass die Schülerinnen und Schüler beim Lösen der Aufgaben mehrheitlich über intendierte Konzepte nachdenken, kann dies als Hinweis für kognitive Validität aus dem Bereich (I) gedeutet werden. Im Rahmen vorliegender Arbeit wird zudem untersucht, inwiefern qualitativ hochwertigere Schüleraussagen mit einer besseren Lösung der Aufgaben einhergehen (Hinweise für kognitive Validität aus dem Bereich (II)).

Deshalb schätzten die Expertinnen und Experten, falls eine kategorisierte Schüleraussage einem intendierten Konzept entsprach, auch die Qualität (bzw. das Niveau) des gezeigten Konzepts ein.

In den folgenden Ausführungen wird zuerst erläutert, welche Expertinnen und Experten am Rating teilnahmen und wie das Rating durchgeführt wurde. Im Anschluss wird der Aufbau des Ratings beschrieben und mit Ausschnitten aus dem Rating illustriert.

8.2.1 Expertinnen und Experten und Durchführung des Ratings

Am Expertenrating nahmen insgesamt acht Fachdidaktikerinnen und Fachdidaktiker teil: Drei Expertinnen und Experten aus dem Bereich Physik-, zwei Expertinnen und Experten aus dem Bereich Chemie-, zwei Expertinnen und Experten aus dem Bereich Biologie- und eine Expertin aus dem Bereich Naturwissenschaftsdidaktik. Drei Expertinnen und Experten habe eine Professur im Bereich einer naturwissenschaftsdidaktischen Disziplin, zwei Expertinnen und Experten haben in diesem Bereich promoviert und drei Expertinnen und Experten arbeiten in einer naturwissenschaftsdidaktischen Forschungsgruppe. Zwei Expertinnen und Experten sind aus Deutschland und sechs aus der Schweiz (von drei verschiedenen Hochschulen). Vier der acht Expertinnen und Experten kannten das Projekt ExKoNawi bereits ausführlicher, da sie bei der aktuellen Studie (vgl. Gesamtvalidierungsstudie, Unterkapitel 6.1) oder in vorherigen Pilotstudien mitwirkten.

Das Expertenrating wurde anhand eines pdf-Formulars (vgl. Anhang, Teil C) durchgeführt, welches die Expertinnen und Experten ausfüllten und per E-Mail retournierten. Im pdf-Formular schätzten die Expertinnen und Experten 85 kategorisierte Schüleraussagen diesbezüglich ein, inwiefern diese den intendierten Konzepten entsprechen. Der geschätzte Zeitaufwand betrug 60 Minuten.

8.2.2 Aufbau des Expertenratings

Das Expertenrating ist im Anhang in Teil C aufgeführt. In den folgenden Ausführungen wird der Aufbau des Ratings beschrieben und mit Ausschnitten aus dem Rating illustriert.

Die erste Seite des Ratings ist eine einführende Seite, auf welcher die Ziele der vorliegenden Studie beschrieben und die Intension des Ratings erläutert werden. Danach folgt ein *aufgabenspezifischer Teil*. Beim aufgabenspezifischen Teil schätzten die Expertinnen und Experten einerseits ein, inwiefern das Durchführen von Messwiederholungen und das Messen mit einer Menge bei den jeweiligen Aufgaben des Problemtyps «Messen» naheliegend sind. Daraufhin

beurteilten die Expertinnen und Experten kategorisierte Schüleraussagen im Bereich Mengenvergrösserung und im Bereich der Wahl des Messinstruments im Hinblick darauf, inwiefern diese den intendierten Konzepten entsprechen. Die kategorisierten Schüleraussagen im Bereich Mengenvergrösserung und der Wahl des Messinstruments sind sehr aufgabenspezifisch und wurden darum in diesem Teil des Ratings aufgeführt. Nach dem aufgabenspezifischen Teil des Ratings folgt ein *allgemeinerer Teil*. Im allgemeineren Teil schätzten die Expertinnen und Experten kategorisierte Schüleraussagen im Bereich Messwiederholung ein, indem sie beurteilten, inwiefern diese Aussagen den intendierten Konzepten entsprechen. Die kategorisierten Schüleraussagen im Bereich Messwiederholung sind weniger aufgabenspezifisch und konnten somit in einem allgemeineren Teil des Ratings für alle Aufgaben des Problemtyps «Messen» zusammengefasst werden.

In den folgenden Ausführungen wird zuerst der aufgabenspezifische Teil des Ratings am Beispiel der Bohnenaufgabe vorgestellt und mit Ausschnitten aus dem Rating illustriert. Hierbei werden auch die Skalen zur Einschätzung vorgestellt. Danach wird der allgemeinere Teil des Ratings erläutert.

Aufgabenspezifischer Teil des Ratings

Der aufgabenspezifische Teil des Ratings umfasst pro Aufgabe einen Teil, bei dem eingeschätzt wird, inwiefern die intendierten Konzepte bei den jeweiligen Aufgaben nahliegend sind (naheliegende Konzepte). Dieser Teil ist bei allen Aufgaben analog strukturiert: Zuerst wird die Aufgabe eingeführt und daraufhin schätzten die Expertinnen und Experten ein, inwiefern das Durchführen von Messwiederholungen und das Messen mit einer Menge bei der Aufgabe naheliegend sind. Hierfür stand den Expertinnen und Experten eine vierstufige Skala (1: nicht naheliegend, 2: eher nicht naheliegend, 3: eher naheliegend, 4: naheliegend) zur Verfügung und am Ende der Einschätzungen konnten die Expertinnen und Experten in einem offenen Antwortfenster weitere Rückmeldungen zur Einschätzung oder Aufgabe geben (vgl. Abb. 15). Hier konnten sie beispielsweise angeben, ob sie bei der Aufgabe Aspekte erkennen, welche die kognitive Validität beeinträchtigen könnten.

Bohnenaufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, wie schwer eine einzelne Bohne ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten
- Waage A: 10 g-Skalierung.
- Waage B: 2 g-Skalierung.
- Die Masse einer getrockneten Bohne beträgt circa 0.3 g bis 1 g.

	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Messwiederholung bedeutet bei dieser Aufgabe mehrmals hintereinander mit einer bestimmten Anzahl Bohnen zu messen und dabei die Bohnen auszutauschen.	○	○	○	○
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (z. B. 20 Bohnen auf einmal) zu messen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Die Masse einer einzelnen Bohne kann mit den gegebenen Messinstrumenten kaum ermittelt werden und wird sehr ungenau. Wenn mit einer Menge gemessen wird (z. B. 20 Bohnen auf einmal), wird das Ergebnis genauer und verschiedene Bohnengrössen werden berücksichtigt.	○	○	○	○

Möglichkeit für Rückmeldungen zur Einschätzung:

Abbildung 15: Ausschnitt aus dem Expertenrating bei der Bohnenaufgabe im Bereich der Einschätzung, inwiefern die Konzepte im Bereich Messwiederholung und Mengenvergrösserung naheliegend sind.

Nach der Einschätzung, inwiefern die Konzepte bei den jeweiligen Aufgaben naheliegend sind, folgt bei allen Aufgaben ein weiterer Teil, bei welchem die Expertinnen und Experten kategorisierte Schüleraussagen im Bereich Mengenvergrösserung einschätzten, welche meistens sehr aufgabenspezifisch sind (Konzepte im Bereich Mengenvergrösserung). Auch hier ist der Aufbau bei allen Aufgaben des Problemtyps «Messen» analog. Auf einer einführenden Seite wird das intendierte Konzept beschrieben, zum Beispiel bei der Bohnenaufgabe: *«Wenn mit einer grossen Menge gemessen wird (z. B. 10 Bohnen auf einmal), dann erhöht dies die Messgenauigkeit»*. Daraufhin folgen weitere Hinweise zur Einschätzung. Ein solcher Hinweis ist beispielsweise, dass bei der Einschätzung, ob ein intendiertes Konzept aktiviert wurde, nur die Frage zählt, ob ein Bezug zum Konzept hergestellt wurde und nicht, ob das Konzept auch richtig angewandt wurde, denn bei den Hinweisen für kognitiven Validität aus dem Bereich (I) geht es nur um die Aktivierung der Konzepte und nicht um deren Richtigkeit. Danach wird die Einschätzung anhand von Beispielen erklärt (vgl. Abb. 16).

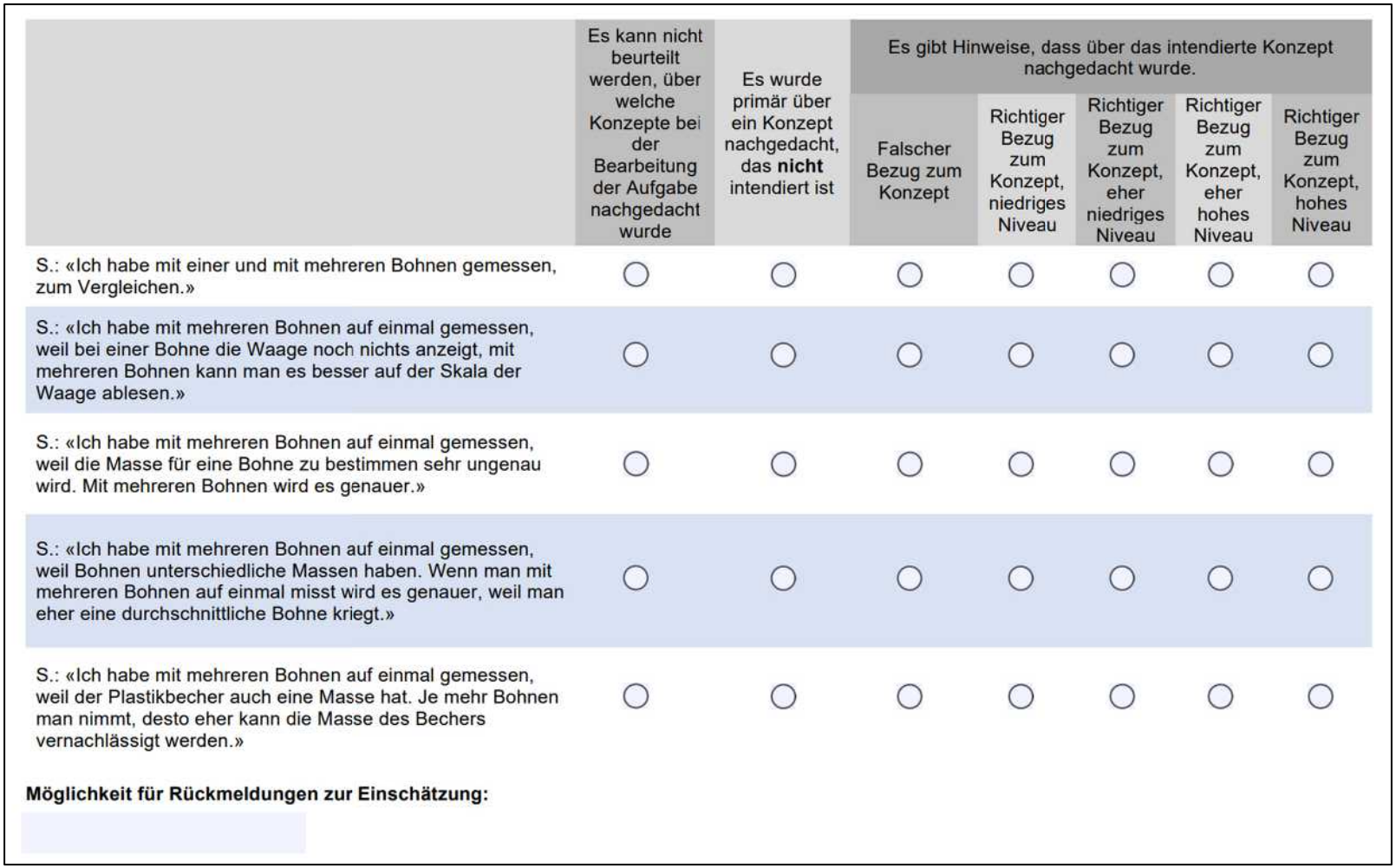
Zu prüfen wäre also bei der Bohnenaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der Anzahl gemessener Bohnen und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe 20 Bohnen auf einmal genommen, damit die Messung genauer wird» oder «Ich habe 10 Bohnen auf einmal genommen, weil ich es so genauer bei der Waage ablesen kann») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nicht über die Anzahl Bohnen nachgedacht, weil diese keinen Einfluss auf die Genauigkeit hat»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe 5 Bohnen auf einmal genommen, weil ich nicht so viele Bohnen abzählen wollte», weil hier die Auswahl der Anzahl Bohnen vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Abbildung 16: Ausschnitt aus dem Expertenrating zur Illustration, wie die Einschätzung von kategorisierten Schüleraussagen anhand von Beispielen erklärt wird. Hier bei der Bohnenaufgabe im Bereich Mengenvergrösserung. Die Abkürzung S.-Aussagen steht für Schüleraussagen.

Der Ausschnitt aus dem Expertenrating in Abbildung 16 zeigt, dass es darum geht einzuschätzen, ob das intendierte Konzept aktiviert wurde und somit über den Zusammenhang zwischen dem Messen mit einer Menge und der Messgenauigkeit nachgedacht wurde. Dabei muss das gezeigte Konzept nicht zwingend korrekt sein (z. B. *«Ich habe nicht über die Anzahl Bohnen nachgedacht, weil diese keinen Einfluss auf die Genauigkeit hat»*): Sobald ersichtlich wird, dass anscheinend über den Zusammenhang mit der Messgenauigkeit nachgedacht wurde, dann ist dies als die Aktivierung eines intendierten Konzepts zu deuten, wenngleich in diesem Fall mit falschem Bezug. Nach der einführenden Seite mit den

Hinweisen zum intendierten Konzept im Bereich Mengenvergrösserung und den Beispielen zur Einschätzung werden die Expertinnen und Experten aufgefordert, kategorisierte Schüleraussagen in diesem Bereich einzuschätzen. Hierfür stand den Expertinnen und Experten eine 7-teilige Skala zur Verfügung (vgl. Abb. 17). Zudem hatten sie auch hier die Möglichkeit für weitere Rückmeldungen in einem offenen Antwortfenster.

Nach den Einschätzungen zu den kategorisierten Schüleraussagen im Bereich Mengenvergrösserung folgt im Rating ein Teil, bei welchem es um Konzepte im Bereich der Wahl des Messinstruments geht, welche auch sehr aufgabenspezifisch sind. Dieser Teil ist gleich strukturiert wie der Teil zu den Konzepten im Bereich Mengenvergrösserung und ist für alle Aufgaben des Problemtyps «Messen» analog aufgebaut: Auf einer einführenden Seite wird das intendierte Konzept erläutert und mit Beispielen die Einschätzung illustriert. Daraufhin schätzten die Expertinnen und Experten ebenfalls anhand einer 7-teiligen Skala (vgl. auch Abb. 17) kategorisierte Schüleraussagen im Bereich der Wahl des Messinstruments ein. Am Ende der Einschätzungen stand wiederum ein offenes Antwortfenster zur Verfügung, in welchem die Expertinnen und Experten Rückmeldungen zur Einschätzung oder Aufgabe geben konnten.



	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe mit einer und mit mehreren Bohnen gemessen, zum Vergleichen.»	○	○	○	○	○	○	○
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil bei einer Bohne die Waage noch nichts anzeigt, mit mehreren Bohnen kann man es besser auf der Skala der Waage ablesen.»	○	○	○	○	○	○	○
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil die Masse für eine Bohne zu bestimmen sehr ungenau wird. Mit mehreren Bohnen wird es genauer.»	○	○	○	○	○	○	○
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil Bohnen unterschiedliche Massen haben. Wenn man mit mehreren Bohnen auf einmal misst wird es genauer, weil man eher eine durchschnittliche Bohne kriegt.»	○	○	○	○	○	○	○
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil der Plastikbecher auch eine Masse hat. Je mehr Bohnen man nimmt, desto eher kann die Masse des Bechers vernachlässigt werden.»	○	○	○	○	○	○	○

Möglichkeit für Rückmeldungen zur Einschätzung:

Abbildung 17: Ausschnitt aus dem Expertenrating im Bereich der Einschätzung von kategorisierten Schüleraussagen zur Mengenvergrösserung am Beispiel der Bohnenaufgabe.

Allgemeinerer Teil des Ratings: Konzepte im Bereich Messwiederholung

Im allgemeineren Teil des Ratings schätzten die Expertinnen und Experten kategorisierte Schüleraussagen im Bereich Messwiederholung im Hinblick darauf ein, inwiefern diese den intendierten Konzepten entsprechen. Die kategorisierten Schüleraussagen im Bereich Messwiederholung sind weniger aufgabenspezifisch und konnten somit für alle Aufgaben zusammengefasst und so den Expertinnen und Experten zur Einschätzung vorgelegt werden. Analog zu den Erläuterungen des aufgabenspezifischen Teils des Ratings wird auch bei diesem Teil auf einer einführenden Seite das intendierte Konzept beschrieben (vgl. Abb. 18).

Intendiertes Konzept

Bei den Aufgaben des Problemtyps «Messen mit vorgegebenen Instrumenten» (Ahorn-, Bohnen-, Faden-, Filzstift-, Münzen- und Pulveraufgabe) geht es auch um das Konzept: Wenn Messungen wiederholt werden, dann erhöht dies die Messgenauigkeit. Die Messgenauigkeit kann erhöht werden, indem z. B. ein Mittelwert als Resultat berechnet wird; ein Intervall angegeben wird (also gesagt wird in welchem Bereich die zu messende Grösse liegt); ein mittlerer Wert ausgewählt wird (Median) oder nochmals nachgemessen wird, als eine Art Bestätigung des vorherigen Ergebnisses. Diese Gründe für das Durchführen von Messwiederholungen befinden sich auf unterschiedlichen Niveaus: Während das Berechnen eines Mittelwerts eher einem hohen Niveau entspricht, entspricht das Wiederholen der Messung zur Bestätigung eher einem niedrigen Niveau.

Abbildung 18: Ausschnitt aus dem Expertenrating zur Illustration, wie das intendierte Konzept beschrieben wird. Hier am Beispiel des intendierten Konzepts im Bereich Messwiederholung.

Auf der einführenden Seite wird zudem die Einschätzung anhand von Beispielen illustriert. Ein Beispiel für ein intendiertes Konzept mit richtigem Bezug ist: *«Ich habe Messungen wiederholt, um ein genaueres Ergebnis zu erhalten»*. Hierbei handelt es sich um einen Bezug zu dem intendierten Konzept, da anscheinend über den Zusammenhang zwischen Messwiederholungen und der Messgenauigkeit nachgedacht wurde. Dabei zeigt der Schüler respektive die Schülerin einen richtigen Bezug zum Konzept (‘wenn man mehrmals misst, wird es genauer’). Ein Beispiel für ein intendiertes Konzept mit falschem Bezug ist: *«Es spielt keine Rolle, ob man einmal oder mehrmals misst, wenn das Experiment genau durchgeführt wurde»*. Hierbei handelt es sich auch um ein intendiertes Konzept, da anscheinend auch hier über den Zusammenhang mit der Genauigkeit nachgedacht wurde. Dabei zeigt der Schüler respektive die Schülerin jedoch einen falschen Bezug zum Konzept (‘einmal messen genügt’). Ein Beispiel für ein nicht intendiertes Konzept ist: *«Ich habe mehrmals gemessen, weil ich ausreichend Zeit hatte»*. Hierbei handelt es sich um ein nicht intendiertes Konzept, da

anscheinend nicht die Genauigkeit, sondern ein anderer Grund ('Zeit') handlungsleitend für das Durchführen von Messwiederholungen war. Nach der einführenden Seite mit den Erläuterungen zum intendierten Konzept und den Beispielen zur Einschätzung wurden den Expertinnen und Experten kategorisierte Schüleraussagen im Bereich Messwiederholung zur Einschätzung vorgelegt. Hierfür stand den Expertinnen und Experten wiederum eine 7-teilige Skala zur Verfügung (vgl. auch Abb. 17) und am Ende der Einschätzungen konnten sie wieder Rückmeldungen in einem offenen Antwortfenster geben.

8.3 Auswertung Expertenrating

Die Expertinnen und Experten schätzten unter anderem anhand einer 7-teiligen Skala (vgl. Unterkapitel 8.2.2) ein, inwiefern die kategorisierten Schüleraussagen den intendierten Konzepten entsprechen. Für die Auswertung des Ratings wurden die Kategorien, die Auskunft bezüglich der Niveaus der gezeigten Konzepte geben, zusammengefasst. Dieser Schritt schien erforderlich, da eine erste Sichtung der Ergebnisse des Ratings zeigte, dass unterschiedliche Einschätzungen der Expertinnen und Experten vor allem auch bei der Beurteilung der Niveaus entstanden. Deswegen und weil eine geringere Aufschlüsselung der Niveaus für die Auswertung ausreichend ist, wurden die Niveaus für die weiteren Analysen zusammengenfasst. Somit wurden für die Auswertung nur *fünf* Differenzierungen unterschieden:

- Es kann nicht beurteilt werden, über welche Konzepte nachgedacht wurde. (Code 11)
- Es wurde primär über ein Konzept nachgedacht, das nicht intendiert ist. (Code 22)
- Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde. Falscher Bezug zum Konzept. (Code 1)
- Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde. Richtiger Bezug zum Konzept, niedriges bis eher niedriges Niveau. (Code 2)
- Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde. Richtiger Bezug zum Konzept, eher hohes bis hohes Niveau. (Code 3)

Insgesamt schätzten die acht Expertinnen und Experten 85 kategorisierte Schüleraussagen diesbezüglich ein, inwiefern die Aussagen den intendierten Konzepten entsprechen. Die Beurteilerübereinstimmung zwischen den Expertinnen und Experten bezüglich der fünf Differenzierungen (Code 11, Code 22 und Code 1 bis 3) wurde mit Hilfe des statistischen Masses Fleiss Kappa geprüft, da in die Prüfung das Urteil von mehr als zwei Personen einfliesst. Die Prüfung ergab

gemäss Landis und Koch (1977) eine moderate und somit akzeptable Übereinstimmung (Fleiss $k = .48$, $p \leq .001$).

Um Hinweise für kognitive Validität aus den Bereichen (I) bis (III) (vgl. Unterkapitel 4.2) zu finden und somit zu beurteilen, inwiefern die Aufgaben des Problemtyps «Messen» kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen, ergeben sich für die Auswertung des Ratings mehrere Auswertungsschritte:

(1) Um Hinweise für kognitive Validität aus dem Bereich (I) zu finden muss analysiert werden, inwiefern die Lernenden beim Lösen der Aufgaben mehrheitlich über die intendierten Konzepte nachdenken. Da die Einschätzungen der acht Expertinnen und Experten im Hinblick darauf, inwiefern eine kategorisierte Schüleraussage einem intendierten Konzept entspricht, in manchen Fällen unterschiedlich sind, muss eine Einschätzung für die weiteren Analysen festgelegt werden. Wie diesbezüglich vorgegangen wurde, wird in Unterkapitel 8.3.1 erläutert.

(2) Um Hinweise für kognitive Validität aus dem Bereich (II) zu finden, muss analysiert werden, inwiefern qualitativ hochwertigere Denkprozesse mit einer besseren Lösung der Aufgabe einhergehen. Hierfür muss anhand der Einschätzungen der Expertinnen und Experten bezüglich des Niveaus der gezeigten Konzepte ein Score gebildet werden, der unterschiedliche Ausprägungen hinsichtlich der Qualität aufzeigen kann. Dafür wurde anhand der Einschätzungen der Expertinnen und Experten ein Q.i.K.-Score (*Q*ualität der *i*ntendierten *K*onzepte-Score) gebildet. In Unterkapitel 8.3.2 wird die Vorgehensweise zur Bildung des Q.i.K.-Scores beschrieben.

(3) Um Hinweise für kognitive Validität aus dem Bereich (III) zu finden, muss analysiert werden, inwiefern das Durchführen von Messwiederholungen und das Messen mit einer Menge von den Expertinnen und Experten bei den Aufgaben des Problemtyps «Messen» als naheliegend eingeschätzt wurden (Plausibilität) und ob die Expertinnen und Experten bei den Aufgaben Aspekte erkennen, welche die kognitive Validität beeinträchtigen könnten. Wie diesbezüglich vorgegangen wurde, wird in Unterkapitel 8.3.3 beschrieben.

8.3.1 Festlegen auf eine Einschätzung für die weitere Auswertung

Die Einschätzungen der Expertinnen und Experten diesbezüglich, inwiefern eine kategorisierte Schüleraussage einem intendierten Konzept entspricht, waren in manchem Fällen unterschiedlich. Für die weitere Auswertung war es somit notwendig, für jede kategorisierte Schüleraussage einen Code (Code 11, Code 22 oder Code 1 bis 3) für die weiteren Analysen festzulegen. Wenn sich die

Expertinnen und Experten mit ihrer Einschätzung einig waren, wurde die Einschätzung so übernommen, wie diese von den Expertinnen und Experten vorgenommen wurde (vgl. Tab. 14, Zeile 1). In den Fällen, in denen die Expertinnen und Experten kategorisierte Schüleraussagen unterschiedlich einschätzten, wurde meistens die mehrheitliche Einschätzung für die weitere Auswertung übernommen (vgl. Tab. 14, Zeile 2 und 3). Unterschiedliche Einschätzungen bezüglich «Es kann nicht beurteilt werden, über welche Konzepte nachgedacht wurde» (Code 11) und «Es wurde primär über ein Konzept nachgedacht, das nicht intendiert ist» (Code 22) wurden erneut mit einer Fachdidaktikerin und einem Fachdidaktiker, welche mit den Aufgabenstellungen und Intensionen der Aufgaben vertraut sind, besprochen und dadurch konnte ein Code für die weitere Auswertung festgelegt werden (vgl. Tab. 14, Zeile 4 und 5)[36]. Dieser Schritt erschien erforderlich, da unterschiedliche Einschätzungen bezüglich Code 11 und 22 einen Einfluss auf die Suche nach Hinweisen für kognitive Validität aus dem Bereich (I) haben können: Während Code 11 eine nicht bedrohende Einschränkung für die kognitive Validität darstellt, zeigt Code 22 auf, dass anscheinend über nicht intendierte Konzepte nachgedacht wurde. Somit ist Code 22 als Hinweis zu deuten, der gegen die kognitive Validität spricht. Zudem kam es öfters zu unterschiedlichen Einschätzungen bezüglich des Niveaus der kategorisierten Schüleraussagen. Diese unterschiedlichen Einschätzungen können einen Einfluss auf die Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) haben, denn hier wird untersucht, inwiefern qualitativ höhere Denkprozesse (höheres Niveau) mit einer besseren Lösung der Aufgabe einhergehen. Deshalb wurde der Einfluss der gewählten Einschätzung für die weitere Auswertung geprüft, indem bei allen Einschätzungen, bei denen die Expertinnen und Experten das Niveau unterschiedlich einschätzten, für alle diese Aussagen einmal die 'strengere'[37] und einmal die 'weniger strenge'[38] Einschätzung für die Auswertung verwendet wurde. Dabei konnte festgestellt werden, dass der Einfluss der gewählten Einschätzung gering ist (Tendenz des Zusammenhangs bleibt gleich und Korrelationskoeffizienten unterscheiden sich nur geringfügig) und somit angenommen werden kann, dass das Festlegen auf eine Einschätzung für die weitere Auswertung nur einen kleinen Einfluss auf

[36] Zu unterschiedlichen Einschätzungen bezüglich Code 11 und 22 kam es öfters bei kategorisierten Schüleraussagen, bei welchen anhand der Aufgabenstellung argumentiert wurde (z. B. *«Ich habe mit einer Pulthöhe gemessen, da es so in der Aufgabe steht»*). Einige Expertinnen und Experten schätzten diese Aussagen mit Code 11 und andere mit Code 22 ein. Nach der weiteren Austauschrunde mit einer Fachdidaktikerin und einem Fachdidaktiker wurde bei diesen Aussagen Code 11 («Es kann nicht beurteilt werden, über welche Konzepte nachgedacht wurde») für die fortführende Auswertung festgelegt.

[37] 'strengere' Einschätzung bedeutet, dass das tiefer eingeschätzte Niveau gewählt wurde.

[38] 'weniger strengere' Einschätzung bedeutet, dass das höher eingeschätzte Niveau gewählt wurde.

die Ergebnisse und keinen Einfluss auf die Interpretation der Ergebnisse hat. Bei den kategorisierten Schüleraussagen, bei denen das Niveau unterschiedlich eingeschätzt wurde, wurde daraufhin die ‘strengere’ Einschätzung für die weitere Auswertung übernommen (vgl. Tab. 14, Zeile 4 und 5). Dies wurde so festgelegt, da diese Einschätzung mit der Einschätzung der Autorin vorliegender Arbeit übereinstimmt. In wenigen Fällen waren die Einschätzungen der Expertinnen und Experten sehr heterogen. Hier wurde von der Autorin eine Einschätzung für die weitere Auswertung festgelegt (vgl. Tab. 14, Zeile 6).

Tabelle 14: Übereinstimmungen der Einschätzungen der Expertinnen und Experten und Erläuterungen dazu, inwiefern eine Einschätzung für die weitere Auswertung festgelegt wurde.

Zeilen	**Übereinstimmung**	**Festlegen der Einschätzung für die weitere Auswertung**
1	Bei 21 von 85 kategorisierten Schüleraussagen (≙ 25 % der Aussagen) haben alle Expertinnen und Experten die Aussage gleich eingeschätzt.	Die Einschätzung der Expertinnen und Experten wurde übernommen.
2	Bei 13 von 85 Aussagen (≙ 15 % der Aussagen) haben 7 von 8 Expertinnen und Experten die Aussagen gleich eingeschätzt.	Für die Auswertung wurde die mehrheitliche Einschätzung übernommen.
3	Bei 16 von 85 Aussagen (≙ 19 % der Aussagen) haben 6 von 8 Expertinnen und Experten die Aussagen gleich eingeschätzt.	
4	Bei 12 von 85 Aussagen (≙ 14 % der Aussagen) haben 5 von 8 Expertinnen und Experten die Aussagen gleich eingeschätzt. Unterschiedliche Einschätzungen entstanden v.a. bei Code 11 und 22 und bezüglich des eingeschätzten Niveaus (Code 2 und 3).	<u>Unterschiedliche Einschätzungen bezüglich Code 11 und 22:</u> Erneute Besprechung mit einer Fachdidaktikerin und einem Fachdidaktiker und dadurch Festlegen auf eine Einschätzung für die weitere Auswertung. <u>Unterschiedliche Einschätzung bezüglich des Niveaus:</u> Der Einfluss der festgelegten Einschätzung wurde geprüft. Der Einfluss erwies sich als gering. Die ‘strengere‘ Einschätzung wurde für die Auswertung übernommen, analog zur Einschätzung der Autorin vorliegender Arbeit.
5	Bei 19 von 85 Aussagen (≙ 22 % der Aussagen) hat die Hälfte der Expertinnen und Experten die Aussagen gleich eingeschätzt. Unterschiedliche Einschätzungen entstanden v.a. bei Code 11 und 22 und bezüglich des eingeschätzten Niveaus (Code 2 und 3).	

6	Bei 4 von 85 Aussagen (≙ 5 % der Aussagen) war die Einschätzung der Expertinnen und Experten sehr heterogen.	Hier wurde für die Auswertung eine Einschätzung von der Autorin festgelegt.

8.3.2 Bildung des Q.i.K.-Scores

Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) wird untersucht, ob Schülerinnen und Schüler, die gemäss Einschätzungen der Expertinnen und Experten ein hohes Niveau bei den intendierten Konzepten zeigen, auch eine bessere Lösung der Aufgabe erzielen. Wenn ein solcher Zusammenhang gegeben ist, kann dies als Hinweis für kognitive Validität aus dem Bereich (II) gedeutet werden. Somit wurde für die Auswertung anhand des im Rating eingeschätzten Niveaus ein Q.i.K.-Score (*Q*ualität der *i*ntendierten *K*onzepte-Score) gebildet, der unterschiedliche Ausprägungen bezüglich der Qualität der gezeigten Konzepte aufzeigt. Zur Bildung des Q.i.K.-Scores wurden die festgelegten Einschätzungen von Unterkapitel 8.3.1 verwendet. Der Q.i.K.-Score besteht aus Punkten im Bereich Messwiederholung (MW), Mengenvergrösserung (Messen mit einer (grossen) Menge, GM) und Wahl des Messinstruments (MI) (vgl. Tab. 15).

Tabelle 15: Beschreibung der Vorgehensweise zur Bildung des Q.i.K.-Scores (Qualität der intendierten Konzepte-Score).

Zeilen	Intendierte Konzepte (i.K.)			Festgelegte Einschätzung	Punkte Q.i.K.-Score
	Durch das Durchführen von **MW** kann die Messgenauigkeit erhöht werden.	Durch das Messen mit einer **GM** kann die Messgenauigkeit erhöht werden.	Für eine präzise Messung sollte immer ein möglichst genaues **MI** verwendet werden.		
	Beispiele von kategorisierten Schüleraussagen				
1	*Einmal messen genügt, wenn man das Experiment genau durchführt.*	*Es ist besser mit einem Faden aufs Mal zu messen. Mit doppeltem Faden wird es ungenau.*	*Thermometer A sieht genauer aus, da es grösser ist.*	**Code 1:** i.K., falscher Bezug zum Konzept	1

2	*Mehrmals messen ist besser, weil man so sein Ergebnis bestätigen kann.*	*Ich habe mit mehreren Bohnen gemessen, weil man es so besser auf der Skala der Waage ablesen kann.*	*Ich habe Stoppuhr B genommen, weil diese genauer ist. Ohne weitere Begründung.*	**Code 2:** i.K., richtiger Bezug zum Konzept, eher niedriges Niveau	2
3	*Mehrmals messen ist besser, weil man so ein genaueres Ergebnis kriegt und einen Mittelwert berechnen kann.*	*Ich habe mit mehreren Bohnen aufs Mal gemessen, weil es so genauer wird und man unterschiedliche Bohnengrössen berücksichtigen kann.*	*Ich habe mit Federwaage A gemessen, da diese genauer ist. Mit Verweis auf die feinere Skala.*	**Code 3:** i.K., richtiger Bezug zum Konzept, eher hohes Niveau	3
4	*Mehrmals messen ist besser, ohne weitere Begründung.*	*Es ist besser mit 1 Tütchen Pulver auf einmal zu messen, da es so in der Aufgabe steht.*	*Ich habe mit Thermometer A gemessen, da Thermometer B nicht richtig funktionierte.*	**Code 11:** Es kann nicht beurteilt werden, über welche Konzepte nachgedacht wurde	0
5	*Einmal messen genügt, weil man hat keine Zeit für MW.*	*Ich habe mit mehr als 1 Tütchen Pulver gemessen, aus Neugierde.*	*Ich habe mit Federwaage B gemessen, weil ich Angst hatte, dass Federwaage A kaputt geht.*	**Code 22:** Kein i.K.	0
Zeilen	Verbesserungsvorschläge im Bereich der i.K. - Äussert **MW** als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit. - Äussert das Messen mit einer **GM** als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit. - Äussert die Wahl eines genaueren **MI** als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit.				Punkte Q.i.K.-Score
6	Keine Äusserung zu MW / Messen mit einer GM / Wahl eines genaueren MI als Verbesserungsvorschlag.				+ 0
7	Denkt beim Lösen der Aufgabe anscheinend bereits über den Zusammenhang zwischen MW / dem Messen mit einer GM / der Wahl des MI und der Messgenauigkeit nach (vgl. Zeilen 1 bis 3). MW bzw. noch mehr MW / Messen mit einer GM bzw. noch grösseren Menge / ein genaueres MI wird als Verbesserungsvorschlag geäussert.				+ 1

8	Denkt beim Lösen der Aufgabe anscheinend nicht über den Zusammenhang zwischen MW / dem Messen mit einer GM / der Wahl des MI und der Messgenauigkeit nach bzw. es wird nicht ersichtlich (vgl. Zeilen 4 und 5). MW / Messen mit einer GM / ein genaueres MI wird aber als Verbesserungsvorschlag geäussert.	+ 2

Anmerkung: MW steht für Messwiederholung, GM steht für (grosse) Menge und MI steht für Messinstrument. Die Beispiele der kategorisierten Schüleraussagen beziehen sich auf die Aufgaben des Problemtyps «Messen» (vgl. Unterkapitel 6.4.1).

Regeln zur Bildung des Q.i.K.-Score (vgl. Tab. 15):

Einerseits gab es Punkte für den Q.i.K.-Score, wenn ersichtlich wurde, dass der Schüler respektive die Schülerin beim Lösen der Aufgabe anscheinend über ein intendiertes Konzept (also über den Zusammenhang von dem Durchführen von Messwiederholungen, dem Messen mit einer Menge oder der Wahl des Messinstruments und der Messgenauigkeit) nachgedacht hat. Dabei gab es für ein intendiertes Konzept mit falschem Bezug (Code 1) einen Punkt, für ein intendiertes Konzept mit richtigem Bezug und eher niedrigem Niveau (Code 2) zwei Punkte und für ein intendiertes Konzept mit richtigem Bezug und eher hohem Niveau (Code 3) drei Punkte. Wenn anhand einer kategorisierten Schüleraussage nicht ersichtlich wurde, über welche Konzepte nachgedacht wurde (Code 11) oder wenn über ein nicht intendiertes Konzept nachgedacht wurde (Code 22), dann gab es keine Punkte für den Q.i.K.-Score. Somit konnten in diesen Bereichen jeweils maximal drei Punkte erzielt werden (vgl. Zeilen 1 bis 5). Andererseits gab es Punkte für den Q.i.K.-Score, wenn auf ein intendiertes Konzept als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit verwiesen wurde (vgl. Zeilen 6 bis 8). Zwar wurde in diesen Fällen das intendierte Konzept nicht zwingend zum Lösen der Aufgabe verwendet (z. B. es wurden keine Messwiederholungen gemacht, da keine Zeit), dennoch wird durch den geäusserten Verbesserungsvorschlag (z. B. «Wenn man mehr Zeit hätte, dann würde ich dreimal messen, da es so genauer wird») ersichtlich, dass der Schüler oder die Schülerin den Zusammenhang zwischen dem Konzept und der Messgenauigkeit erkannt hat. Falls bereits über ein intendiertes Konzept zum Lösen der Aufgabe nachgedacht wurde (vgl. Zeilen 1 bis 3) und das Durchführen von (noch mehr) Messwiederholungen, das Messen mit einer grossen Menge (bzw. noch grösseren Menge) oder ein genaueres Messinstrument als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit genannt wurden, dann gab es hierfür einen zusätzlichen Punkt (vgl. Zeile 7). Falls anscheinend beim Lösen der Aufgabe nicht über ein intendiertes Konzept nachgedacht oder dies anhand der kategorisierten Schüleraussage nicht ersichtlich wurde (vgl. Zeilen 4 und 5), bei den Verbesserungsvorschlägen jedoch auf ein intendiertes Konzept verwiesen wurde, gab es hierfür

zwei zusätzliche Punkte (vgl. Zeile 8). Hier gab es zwei zusätzliche Punkte, damit es sich mit den Punkten derjenigen Schülerinnen und Schülern ausgleicht, die beim intendierten Konzept einen richtigen Bezug zum Konzept von eher niedrigem Niveau zeigten (vgl. Zeile 2). Somit konnten bei den Verbesserungsvorschlägen in den Bereichen Messwiederholung, Messen mit einer Menge und Wahl des Messinstruments jeweils maximal zwei zusätzliche Punkte erzielt werden (vgl. Zeilen 6 bis 8).

Falls ein Schüler oder eine Schülerin in einem Bereich mehrere intendierte Konzepte zeigte (dies ist möglich, weil im Bereich Messwiederholung und Messen mit einer Menge jeweils zwei Oberkategorien für die Beurteilung berücksichtigt wurden), wurde darauf geachtet, ob Messwiederholungen durchgeführt wurden respektive mit einer Menge gemessen wurde. Wenn Messwiederholungen gemacht wurden respektive mit einer Menge gemessen wurde: Für die Bildung des Q.i.K.-Scores wurde das Konzept gewählt, das gemäss Einschätzungen der Expertinnen und Experten als höheres Niveau beurteilt wurde. Falls keine Messwiederholungen gemacht wurden respektive nicht mit einer Menge gemessen wurde, dann wurden die Oberkategorien zur Bildung des Q.i.K.-Scores verwendet, die Auskunft darüber geben, warum bei der Aufgabe nur einmal gemessen wurde (Oberkategorie 7; vgl. Kategoriensystem im Anhang, Teil B) beziehungsweise nicht mit einer Menge gemessen wurde (Oberkategorie 16). Die anderen Oberkategorien («Ist das Durchführen von Messwiederholungen respektive das Messen mit einer Menge *prinzipiell* bei der Aufgabe sinnvoll», vgl. Oberkategorien 4 und 15) wurden als Verbesserungsvorschläge zur Steigerung der Messgenauigkeit gezählt.

Um zu beurteilen, ob ein genaueres Messinstrument (vgl. Zeilen 6 bis 8) und / oder noch mehr Messwiederholungen respektive eine noch grössere Menge (vgl. Zeile 7) als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit genannt wurden, wurden die Interviews an entsprechender Stelle («Kannst du Vorschläge machen, wie man noch genauer messen könnte», vgl. Interviewleitfaden Anhang, Teil A) und die Schülerprotokolle bei entsprechendem Auftrag («Wie könntest du noch genauer messen? Mache Vorschläge und erkläre, wieso man so genauer messen kann», vgl. Unterkapitel 6.4.2) erneut gesichtet und so die zusätzlichen Punkte für die Verbesserungsvorschläge verteilt.

In Tabelle 16 wird die Punktevergabe für die Bildung des Q.i.K.-Scores anhand eines Beispiels bei der Pulveraufgabe illustriert. Dabei werden die Regeln zur Bildung des Q.i.K.-Scores von Tabelle 15 verwendet.

Tabelle 16: Illustration der Punktevergabe für die Bildung des Q.i.K.-Scores (Qualität der intendierten Konzepte-Score) am Beispiel von Antworten einer Schülerin bei der Pulveraufgabe.

		Beispielantworten einer Schülerin bei der Pulveraufgabe und Erklärungen zur Punktevergabe	**Punkte Q.i.K.-Score**
MW	Intendiertes Konzept	*«Ich habe mehrmals gemessen, damit es genauer wird und ich einen Mittelwert berechnen kann».* Die Beispielantwort der Schülerin zeigt, dass sie anscheinend beim Lösen der Aufgabe über den Zusammenhang zwischen der Messgenauigkeit und Messwiederholung nachgedacht hat. Dabei wurde die Antwort der Schülerin als eher hohes Niveau eingestuft. Folglich erhält die Schülerin bei den intendierten Konzepten im Bereich Messwiederholung drei Punkte.	3
	Verbesserungsvorschlag	*«Beim nächsten Mal würde ich noch mehr Messungen machen, damit mein Ergebnis noch genauer wird».* Die Schülerin äussert das Durchführen von noch mehr Messwiederholungen als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit. Somit erhält sie bei den Verbesserungsvorschlägen im Bereich Messwiederholung einen zusätzlichen Punkt, da sie beim Lösen der Aufgabe offenbar bereits über ein intendiertes Konzept nachgedacht hat.	+ 1
GM	Intendiertes Konzept	*«Ich würde nur mit 1 Tütchen Pulver aufs Mal messen, da es so in der Aufgabe steht».* Die Beispielantwort der Schülerin zeigt, dass im Bereich Messen mit einer Menge nicht ersichtlich wird, ob sie über ein intendiertes Konzept nachgedacht hat. Somit erhält die Schülerin in diesem Bereich keine Punkte.	0
	Verbesserungsvorschlag	- Im Bereich Messen mit einer Menge werden keine Verbesserungsvorschläge genannt. Somit erhält die Schülerin hier keine zusätzlichen Punkte.	+ 0

MI	Intendiertes Konzept	*«Ich habe Thermometer A genommen, da B nicht richtig funktionierte».* Die Beispielantwort der Schülerin zeigt, dass im Bereich Wahl des Messinstruments nicht beurteilt werden kann, über welche Konzepte die Schülerin nachgedacht hat. Somit erhält die Schülerin in diesem Bereich keine Punkte.	0
	Verbesserungsvorschlag	*«Man könnte genauer messen, wenn man ein genaueres Thermometer hätte, zum Beispiel ein digitales Thermometer».* Die Schülerin nennt die Wahl eines genaueren Messinstruments als Verbesserungsvorschlag zur Steigerung der Messgenauigkeit. Somit erhält sie beim Verbesserungsvorschlag zwei zusätzliche Punkte, da nicht ersichtlich wurde, ob sie beim Lösen der Aufgabe im Bereich der Wahl des Messinstruments bereits über ein intendiertes Konzept nachgedacht hat.	+ 2

Anmerkung: MW steht für Messwiederholung, GM steht für (grosse) Menge und MI steht für Messinstrument.

Tabelle 16 illustriert die Punktevergabe an einem Beispiel bei der Pulveraufgabe. Es wird ersichtlich, dass die Schülerin beim Q.i.K.-Score total sechs Punkte erzielt hat. Insgesamt sind beim Q.i.K.-Score bei allen Aufgaben des Problemtyps «Messen» zwölf Punkte möglich. In den Bereichen MW, GM und MI können jeweils maximal vier Punkte erreicht werden: Zeigt beim Lösen der Aufgabe ein intendiertes Konzept mit richtigem Bezug und eher hohem Niveau (drei Punkte) und äussert zudem einen Verbesserungsvorschlag im Bereich des intendierten Konzepts (einen zusätzlichen Punkt).

8.3.3 Analyse der Plausibilität der Konzepte und mögliche Aspekte, welche die kognitive Validität beeinträchtigen könnten

Nur wenn die intendierten Konzepte bei den Aufgaben des Problemtyps «Messen» naheliegend sind (Plausibilität), werden diese wahrscheinlich auch seitens der Lernenden aktiviert (Hinweise für kognitive Validität aus dem Bereich (III)). Für die Auswertung der Plausibilität wurde von den Einschätzungen der acht Expertinnen und Experten im Hinblick darauf, inwiefern die Konzepte im Bereich Messwiederholung und Mengenvergrösserung naheliegend sind (Skala, vgl. Unterkapitel 8.2.2), ein Mittelwert gebildet. Zudem wurde im Rahmen der Suche nach Hinweisen für kognitive Validität aus dem Bereich (III) auch untersucht, ob die Expertinnen und Experten Aspekte erkennen, welche die kognitive Validität beeinträchtigen könnten. Für diese Auswertung wurden die Antworten der Expertinnen und Experten in den offenen Antwortfenstern (vgl. Unterkapitel 8.2.2) qualitativ betrachtet und für die weiteren Analysen zusammengetragen.

8.4 Ergebnisse

Validität ist abhängig vom sozialen Kontext und dem Verwendungszweck (vgl. Kane, 2006; Messick, 1995 und Kapitel 4) und kann somit nicht isoliert betrachtet werden, sondern muss im Zusammenhang mit der Anwendung untersucht werden. Bei vorliegender Arbeit war es das Ziel, Hinweise dafür zu generieren, dass die mit dem Testverfahren erhobenen Ergebnisse eine Aussage über die experimentellen Kompetenzen von Jugendlichen des 8. Schuljahres im Sinne einer Standortbestimmung zulassen. Die in der Folge vorgestellten Hinweise sind somit vor diesem Hintergrund zu verstehen.

In den Unterkapiteln 8.4.1 bis 8.4.3 werden die Ergebnisse der Suche nach Hinweisen für kognitive Validität aus den Bereichen (I) bis (III) beschrieben. Im Rahmen vorliegender Arbeit wurde davon ausgegangen, dass die Aufgaben des Problemtyps «Messen» kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen, wenn Hinweise für kognitive Validität gefunden werden können und keine bedrohenden Einschränkungen vorliegen. Dickmann (2016) unterscheidet zwischen bedrohenden und nicht bedrohenden Einschränkungen im Validierungsprozess:

- Bedrohende Einschränkungen zeigen Hinweise auf, die gegen die kognitive Validität sprechen. Ein Beispiel für eine bedrohende Einschränkung ist, wenn ersichtlich wird, dass die Mehrheit der Schülerinnen und Schüler beim Lösen der Aufgabe über ein nicht intendiertes Konzept nachgedacht hat (vgl. Code 22, Unterkapitel 8.3).
- Nicht bedrohende Einschränkungen ergeben sich aufgrund einer eingeschränkten Datenbasis und zeigen auf, dass in diesem Bereich keine Hinweise für kognitive Validität gefunden werden konnten. Nicht bedrohende Einschränkungen sprechen jedoch nicht *gegen* die kognitive Validität und dadurch wird die Validitätsargumentation nicht bedeutsam geschwächt (Dickmann, 2016). Ein Beispiel für eine nicht bedrohende Einschränkung ist, wenn bei der Mehrheit der Schülerinnen und Schüler nicht ersichtlich wird, über welche Konzepte sie beim Lösen der Aufgabe nachgedacht haben (vgl. Code 11, Unterkapitel 8.3).

8.4.1 Ergebnisse der Suche nach Hinweisen für kognitive Validität aus dem Bereich (I)

Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (I) wird geprüft, ob beim Lösen der Aufgaben des Problemtyps «Messen» mehrheitlich über die intendierten Konzepte nachgedacht wurde. Um dies zu prüfen, wurden die festgelegten Einschätzungen von Unterkapitel 8.3.1 verwendet. Beim Zusammentragen der Ergebnisse wurde betrachtet, wie viel Prozent der Schülerinnen und Schüler beim Lösen der Aufgabe anscheinend über ein intendiertes Konzept nachgedacht haben (vgl. Tab. 17, Symbol '✓'). Wenn die Mehrheit der Lernenden über ein intendiertes Konzept beim Lösen der Aufgabe nachgedacht hat, kann dies als Hinweis für kognitive Validität aus dem Bereich (I) gedeutet werden, da somit ersichtlich wird, dass die Aufgabe vermutlich kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich dieses Konzepts zulässt. Dabei geht es lediglich um eine Aktivierung des Konzepts und nicht um dessen Richtigkeit. Somit können sowohl kategorisierte Schüleraussagen mit richtigem Bezug zum Konzept (vgl. Code 2 und 3 in Unterkapitel 8.3; Beispiel einer Aussage: *«Es wurde mehrmals gemessen, damit es genauer wird»*) als auch mit falschem Bezug zum Konzept (vgl. Code 1; Beispiel einer Aussage: *«Einmal messen genügt, wenn das Experiment genau durchgeführt wurde»*) als Hinweis für kognitive Validität aus dem Bereich (I) gedeutet werden, solange ersichtlich wird, dass über das intendierte Konzept (hier: Zusammenhang zwischen Messwiederholungen und der Messgenauigkeit) nachgedacht wurde. Zudem wurde beim Zusammentragen der Ergebnisse auch betrachtet, wie viel Prozent der Schülerinnen und Schüler beim Lösen der Aufgabe über kein intendiertes Konzept nachgedacht haben (vgl. Tab. 17, Symbol *'×'* und Code 22) und bei wie viel Prozent der Schülerinnen und Schüler nicht ersichtlich wird, über welche Konzepte sie beim Lösen der Aufgabe nachdenken (vgl. Tab. 17, Symbol *'○'* und Code 11). Die Ergebnisse hierzu sind in Tabelle 17 zu finden, wobei jeweils die Werte, welche die Mehrheit in einem Bereich darstellen, fett hervorgehoben sind.

Tabelle 17: Ergebnisse der Suche nach Hinweisen für kognitive Validität aus dem Bereich (I) bei den Aufgaben des Problemtyps «Messen»

Aufgaben		Messwiederholung		Mengen-vergrösserung		Wahl Messinstrument	
		Anzahl SuS	Prozent-zahl	Anzahl SuS	Prozent-zahl	Anzahl SuS	Prozent-zahl
Ahorn	✓	**17 SuS**	**94 %**	2 SuS	11 %	**17 SuS**	**94 %**
	○	-	-	**12 SuS**	**67 %**	1 S	6 %
	×	1 S	6 %	4 SuS	22 %	-	-
Bohne	✓	**15 SuS**	**83 %**	**17 SuS**	**94 %**	**14 SuS**	**78 %**
	○	2 SuS	11 %	-	-	-	-
	×	1 S	6 %	1 S	6 %	4 SuS	22 %
Faden	✓	**18 SuS**	**100 %**	-	-	**12 SuS**	**67 %**
	○	-	-	**14 SuS**	**78 %**	-	-
	×	-	-	4 SuS	22 %	6 SuS	33 %
Filzstift	✓	**18 SuS**	**100 %**	4 SuS	22 %	**18 SuS**	**100 %**
	○	-	-	**14 SuS**	**78 %**	-	-
	×	-	-	-	-	-	-
Münze	✓	**16 SuS**	**89 %**	**16 SuS**	**89 %**	**17 SuS**	**94 %**
	○	1 S	6 %	2 SuS	11 %	1 S	6 %
	×	1 S	6 %	-	-	-	-
Pulver	✓	**13 SuS**	**72 %**	1 S	6 %	4 SuS	22 %
	○	4 SuS	22 %	**12 SuS**	**67 %**	2 SuS	11 %
	×	1 S	6 %	5 SuS	28 %	**12 SuS**	**67 %**

Anmerkung: Das Symbol '✓' bedeutet: Es wurde anscheinend über ein intendiertes Konzept beim Lösen der Aufgabe nachgedacht; '○' bedeutet: Es wird nicht ersichtlich, über welche Konzepte nachgedacht wurde und '×' bedeutet: Es wurde offenbar über kein intendiertes Konzept nachgedacht. Die Abkürzung S steht für Schülerin oder Schüler und SuS steht für Schülerinnen und Schüler. Die Prozentzahlen beziehen sich auf eine Stichprobengrösse von 18 SuS pro Aufgabe (vgl. Unterkapitel 6.2) und stellen gerundete Werte dar. Fett hervorgehoben sind jeweils die Werte, welche die Mehrheit in einem Bereich darstellen.

Tabelle 17 zeigt, dass im Bereich Messwiederholung bei allen Aufgaben des Problemtyps «Messen» anscheinend mehrheitlich (bei 72 bis 100 % der Schülerinnen und Schüler) über ein intendiertes Konzept und somit über den Zusammenhang zwischen dem Durchführen von Messwiederholungen und der Messgenauigkeit nachgedacht wurde. Dies kann als Hinweis für kognitive Validität aus dem Bereich (I) gedeutet werden. Zudem denken im Bereich Messwiederholung nur wenige Schülerinnen und Schüler (0 bis 6 % der Lernenden) beim Lösen der Aufgaben offenbar über kein intendiertes Konzept nach.

Beim Messen mit einer Menge (Strategie Mengenvergrösserung) wurde bei der Bohnen- und Münzenaufgabe anscheinend mehrheitlich (94 bzw. 89 % der Lernenden) über ein intendiertes Konzept und somit über den Zusammenhang zwischen dem Messen mit einer Menge und der Messgenauigkeit nachgedacht. Somit können bei diesen Aufgaben im Bereich Mengenvergrösserung Hinweise für kognitive Validität aus dem Bereich (I) gefunden werden. Bei den restlichen Aufgaben (Ahorn, Faden, Filzstift, Pulver) wurde bei der Mehrheit der Lernenden (67 bis 78 % der Schülerinnen und Schüler) nicht ersichtlich, über welche Konzepte sie beim Lösen der Aufgabe nachdenken (z. B. Aussagen wie: *«Es ist besser mit einer Pulthöhe zu messen, da es so in der Aufgabe steht»*). Dass bei der Mehrheit der Lernenden nicht ersichtlich wird, über welche Konzepte sie beim Lösen der Aufgabe nachdenken, stellt jedoch keine bedrohende Einschränkung im Validitätsargumentationsprozess dar. Einige Lernende (0 bis 28 % der Schülerinnen und Schüler) scheinen im Bereich Mengenvergrösserung beim Lösen der Aufgaben auch über kein intendiertes Konzept nachzudenken (z. B. Aussagen wie: *«Ich habe mit einem Faden und zwei Fäden gemessen, zum Vergleichen»*). Da insgesamt aber bei allen Aufgaben entweder offenbar mehrheitlich über ein intendiertes Konzept nachgedacht wurde (Bohne und Münze) oder bei der Mehrheit der Lernenden nicht ersichtlich wurde, über welche Konzepte sie nachdenken (Ahorn, Faden, Filzstift, Pulver), wird dies als nicht bedrohende Einschränkung betrachtet.

Im Bereich der Wahl des Messinstruments wurde bei der Ahorn-, Bohnen-, Faden-, Filzstift- und Münzenaufgabe anscheinend mehrheitlich (67 bis 100 % der Lernenden) über ein intendiertes Konzept und somit über den Zusammenhang zwischen der Wahl des Messinstruments und der Messgenauigkeit nachgedacht. Folglich können hier Hinweise für kognitive Validität aus dem Bereich (I) gefunden werden. Bei diesen Aufgaben gab es auch einige Lernenden (0 bis 33 % der Schülerinnen und Schüler), die anscheinend beim Lösen der Aufgaben über kein intendiertes Konzept nachgedacht haben (z. B. Aussagen wie: *«Ich habe Waage B genommen, da hier die Skala übersichtlicher ist»*). Da jedoch die

Mehrheit der Lernenden offenbar über ein intendiertes Konzept und somit über den Zusammenhang zwischen der Wahl des Messinstruments und der Messgenauigkeit nachgedacht hat, wurde dies als nicht bedrohende Einschränkung betrachtet. Bei der Pulveraufgabe hingegen wurde im Bereich Wahl des Messinstruments anscheinend mehrheitlich (67 % der Schülerinnen und Schüler) über kein intendiertes Konzept nachgedacht (z. B. Aussagen wie: *«Thermometer B wurde gewählt, da mit diesem der Becher mit Wasser weniger umkippt, da dieses kürzer ist»*; vgl. auch Unterkapitel 6.4.1, Abb. 4). Dies stellt eine bedrohende Einschränkung im Validitätsargumentationsprozess dar und zeigt auf, dass die Pulveraufgabe im Bereich der Wahl des Messinstruments keine kognitiv validen Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden zuzulassen scheint, da bei der Mehrheit der Lernenden beim Lösen der Aufgabe anscheinend nicht ein intendiertes Konzept (Messgenauigkeit), sondern andere Aspekte (z. B. Praktikabilität) handlungsleitend waren.

8.4.2 Ergebnisse der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II)

Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) wird untersucht, ob qualitativ höhere Denkprozesse mit einer besseren Lösung der Aufgabe einhergehen. Hierfür wird der Zusammenhang zwischen dem Q.i.K.-Score (*Q*ualität der *i*ntendierten *K*onzepte-Score, vgl. Unterkapitel 8.3.2) und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle betrachtet. Für die Kompetenzdiagnose werden die Ergebnisse der Schülerprotokolle herangezogen, da im Rahmen der Arbeit untersucht werden soll, inwiefern die Aufgaben und Aufträge in den Schülerprotokollen und die daraus resultierenden Ergebnisse der Kompetenzdiagnose kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden zulassen. Beim Betrachten des Zusammenhangs wurden bei den Schülerprotokollen nur die Punkte im Bereich Messwiederholung und Mengenvergrösserung (QS 3, vgl. Unterkapitel 2.3.2) sowie der Wahl des Messinstruments (QS 4) berücksichtigt, da die intendierten Konzepte nur in diesen Bereichen untersucht und zur Bildung des Q.i.K.-Scores verwendet wurden. Beim Untersuchen des Zusammenhangs wurden der Korrelationskoeffizient Kendall-Tau-b und das Konfidenzintervall für r betrachtet. Konfidenzintervalle bezeichnen einen Bereich, in dem man den gesuchten Wert (hier Korrelationskoeffizient r) mit einer grossen Wahrscheinlichkeit (meistens 95 %) erwarten darf. Dabei werden eine untere und obere Grenze des Bereichs ausgewiesen (Bortz & Schuster, 2010). Im Rahmen vorliegender Arbeit wird ein positiver Zusammenhang zwischen dem Q.i.K.-Score und dem Ergebnis der

Kompetenzdiagnose anhand der Schülerprotokolle vermutet. Das bedeutet, dass höhere Denkprozesse (hoher Q.i.K.-Score) mit einer besseren Lösung der Aufgabe (höheres Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle) einhergehen und somit Hinweise für kognitive Validität aus dem Bereich (II) gefunden werden sollten. Als Hinweise für einen positiven Zusammenhang und somit als Hinweise für kognitive Validität aus dem Bereich (II) werden signifikante und positive Korrelationen zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle oder ausschliesslich positive Konfidenzintervalle für *r*, welche den Wert Null nicht miteinschliessen, gedeutet. Dabei wird nicht erwartet, dass der Zusammenhang zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle hoch ausfällt, da: (1) die Auswertungsverfahren zur Bildung des Q.i.K.-Scores und zur Kompetenzdiagnose anhand der Schülerprotokolle sehr unterschiedlich sind (Q.i.K.-Score: Kategoriensystem und Expertenrating; Schülerprotokolle: Kodiermanual) und (2) im Rahmen von FF1 gezeigt werden konnte, dass anscheinend nicht alle Gedanken der Lernenden ausreichend in den Schülerprotokollen festgehalten werden und somit zusätzliche Interviews zu einem genaueren Ergebnis der Kompetenzdiagnose führen (vgl. Unterkapitel 7.2 und 7.3).

In der Folge werden die Ergebnisse der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) auf der Ebene des gesamten Problemtyps «Messen» ($N = 108$) sowie auf der Ebene der einzelnen Aufgaben ($N = 18$) vorgestellt.

Ergebnisse auf der Ebene des gesamten Problemtyps «Messen»

Bei der graphischen Betrachtung des Zusammenhangs zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle für alle Aufgaben des Problemtyps «Messen» ($N = 108$; vgl. Abb. 19) kann festgestellt werden, dass tendenziell eine höhere Einschätzung bei den Niveaus der gezeigten Konzepte (hoher Q.i.K.-Score) auch mit einem höheren Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle einherzugehen und somit ein positiver Zusammenhang zwischen diesen zwei Massen vorzuliegen scheint.

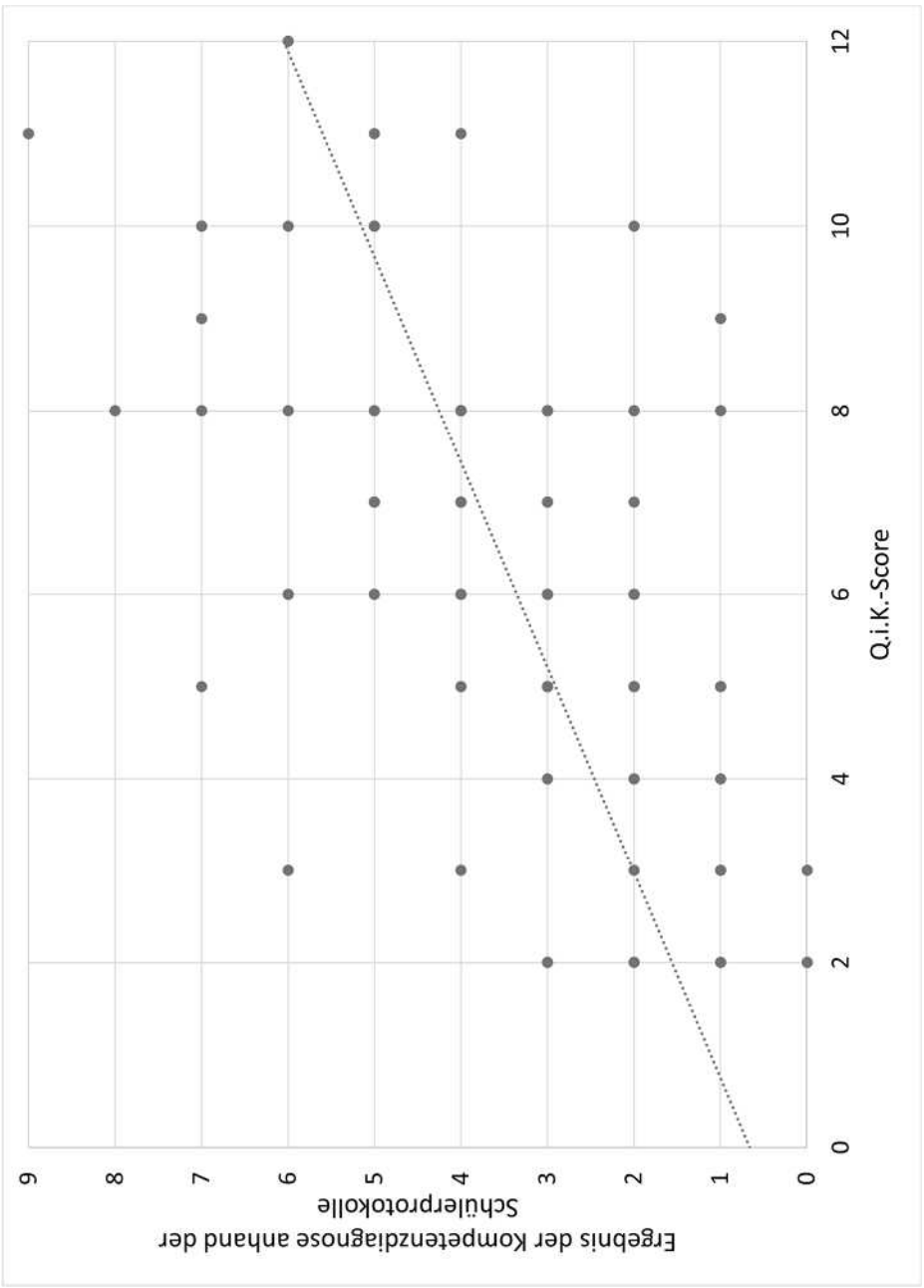


Abbildung 19: Korrelation zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle bei den Aufgaben des Problemtyps «Messen» (N = 108). Anmerkungen: Die Punkte können mehr als einen Schüler bzw. eine Schülerin repräsentieren. Es wurde eine Trendlinie (linear) eingefügt, diese basiert auf allen Datenpunkten. Beim Q.i.K.-Score ist ein Maximum von 12 Punkten möglich (vgl. Unterkapitel 8.3.2). Beim Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle im Bereich von QS 3 und QS 4 ist ein Maximum von 9 Punkten möglich (vgl. Unterkapitel 2.3.2).

Die Berechnung des Korrelationskoeffizienten Kendall-Tau-b zeigt zudem, dass zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle eine signifikante und positive Korrelation ($r = .46$, $p \leq .001$, $N = 108$) vorliegt. Beim Betrachten des Konfidenzintervalls kann festgestellt werden, dass r mit einer Wahrscheinlichkeit von 95 % zwischen .33 und .58 liegt. Somit ist das Konfidenzintervall ausschliesslich positiv und schliesst den Wert Null nicht mit ein. Insgesamt liegen somit bei der Betrachtung aller Aufgaben des Problemtyps «Messen» Hinweise für kognitive Validität aus dem Bereich (II) vor: Höhere Denkprozesse (hoher Q.i.K.-Score) scheinen mit einer besseren Lösung der Aufgabe (höheres Ergebnis bei der Kompetenzdiagnose anhand der Schülerprotokolle) einherzugehen.

Ergebnisse auf der Ebene der einzelnen Aufgaben

In Tabelle 18 sind die Ergebnisse (Korrelationskoeffizient Kendall-Tau-b und Konfidenzintervalle für r) der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) für die Aufgaben des Problemtyps «Messen» aufgeführt.

Tabelle 18: Ergebnisse der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) für die Aufgaben des Problemtyps «Messen».

	Korrelation (Kendall-Tau-b)		**95 % - Konfidenzintervall für *r***	***N***
Ahorn	$r = .57$	$p \leq .01$	untere Grenze: .25 obere Grenze: .82	18
Bohne	$r = .20$	$p = .31$ (n.s.)	untere Grenze: - .17 obere Grenze: .53	18
Faden	$r = .31$	$p = .13$ (n.s.)	untere Grenze: - .07 obere Grenze: .64	18
Filzstift	$r = .28$	$p = .14$ (n.s.)	untere Grenze: - .19 obere Grenze: .73	18
Münze	$r = .33$	$p = .10$ (n.s.)	untere Grenze: - .04 obere Grenze: .65	18
Pulver	$r = .12$	$p = .54$ (n.s.)	untere Grenze: - .28 obere Grenze: .52	18

Beim Betrachten der Korrelationen und Konfidenzintervalle in Tabelle 18 kann festgestellt werden, dass nur bei der Ahornaufgabe eine signifikante und positive Korrelation ($r = .57$, $p \leq .01$, $N = 18$) sowie ein ausschliesslich positives Konfidenzintervall für r beobachtet werden kann (95% - Konfidenzintervall für r; untere Grenze: .25, obere Grenze: .82). Somit liegen bei der Ahornaufgabe Hinweise für kognitive Validität aus dem Bereich (II) vor. Bei den anderen Aufgaben (Bohne, Faden, Filzstift, Münze und Pulver) sind die Korrelationen nicht signifikant und die unteren Grenzen der Konfidenzintervalle sind jeweils negativ.

Somit gibt es bei diesen Aufgaben keine Hinweise auf einen positiven Zusammenhang zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle. Wahrscheinlich entstehen die nicht signifikanten Korrelationen und die nicht ausschliesslich positiven Konfidenzintervalle jedoch aufgrund der geringen Stichprobengrösse (N = 18) auf Aufgabenebene: (1) Bei kleinen Stichproben werden kleine Effekte (hier geringe Korrelationskoeffizienten) oft nicht signifikant (vgl. Bortz & Schuster, 2010) und (2) bei kleinen Stichproben fallen einzelne Ausreisser besonders stark ins Gewicht, was dazu führen kann, dass das Konfidenzintervall 'breiter' wird und somit nicht mehr ausschliesslich positiv ist. Um zu prüfen, ob die nicht ausschliesslich positiven Konfidenzintervalle bei der Bohnen-, Faden-, Filzstift-, Münzen- und Pulveraufgabe womöglich auf eine eingeschränkte Datenbasis zurückzuführen sind und somit als nicht bedrohende Einschränkungen im Validitätsargumentationsprozess betrachtet werden können, wurden in einem nächsten Schritt die Daten auf Ausreisser geprüft. Daraufhin wurde Ausreisser ausgeschlossen und die Konfidenzintervalle erneut betrachtet. Dabei konnte nach Ausschluss von Ausreissern bei der Bohnen-, Faden-, Filzstift- und Münzenaufgabe ein ausschliesslich positives Konfidenzintervall, welches den Wert Null nicht miteinschliessen, beobachtet werden. Dies deutet auf einen positiven Zusammenhang zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle hin. Folglich scheinen die in Tabelle 18 aufgeführten Werte bei der Bohnen-, Faden-, Filzstift- und Münzenaufgabe auf die eingeschränkte Datenbasis zurückzuführen zu sein und stellen damit keine bedrohenden Einschränkungen im Validitätsargumentationsprozess dar. Bei der Pulveraufgabe ist jedoch auch nach Ausschluss von Ausreissern das Konfidenzintervall nicht ausschliesslich positiv. Somit gibt es bei der Pulveraufgabe auch nach Ausschluss von Ausreissern keinen Hinweis auf einen positiven Zusammenhang zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle, sodass davon ausgegangen werden muss, dass das Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle die intendierten Konzepte nicht ausreichend

abzubilden vermag. Die in Tabelle 18 aufgeführten Werte bei der Pulveraufgabe sind somit als bedrohende Einschränkung zu betrachten.[39]

8.4.3 Ergebnisse der Suche nach Hinweisen für kognitive Validität aus dem Bereich (III)

Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (III) wurde einerseits geprüft, ob die Expertinnen und Experten die intendierten Konzepte im Bereich Messwiederholung (MW) und Mengenvergrösserung (Messen mit einer (grossen) Menge, GM) als naheliegend einschätzen (Plausibilität). Denn nur wenn die Expertinnen und Experten diese Konzepte als naheliegend einschätzen, werden diese wahrscheinlich auch seitens der Schülerinnen und Schüler aktiviert. Beim Zusammentragen der Ergebnisse bezüglich der Plausibilität der Konzepte wurde von den Einschätzungen der acht Expertinnen und Experten ein Mittelwert gebildet (Skala vgl. Unterkapitel 8.2.2; 1: nicht naheliegend, 2: eher nicht naheliegend, 3: eher naheliegend, 4: naheliegend; Ergebnisse vgl. Tab. 19). Zudem wurde bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (III) auch untersucht, ob die Expertinnen und Experten Aspekte erkennen, welche die kognitive Validität beeinträchtigen könnten. Hierfür wurden die Antworten der Expertinnen und Experten in den offenen Antwortfenster (vgl. Unterkapitel 8.2.2) berücksichtigt und zusammengetragen (vgl. Tab. 19).

[39] Ergänzend wurde geprüft, ob durch Erhöhung des Stichprobenumfangs die Korrelationen bei der Bohnen-, Faden-, Filzstift-, Münzen- und Pulveraufgabe signifikant werden würden. Hierfür wurde eine Poweranalyse durchgeführt (G-Power-Software der Universität Düsseldorf (G*Power 3.1.9.7); post-hoc-Test; zweiseitiger Test; Signifikanzniveau $p < .05$). Es wurde eine Power von mindestens .8 vorausgesetzt, d.h. dass unter diesen Bedingungen ein signifikanter Effekt mit einer Wahrscheinlichkeit von 80 % nachgewiesen werden kann. Die Poweranalysen zeigten, dass bei der Faden-, Filzstift- und Münzenaufgabe durch eine geringe Erhöhung des Stichprobenumfangs (Faden: $N \geq 75$, Filzstift: $N \geq 95$, Münze: $N \geq 70$) signifikante und positive Korrelationen entstehen würden. Dies unterstützt die Annahme, dass bei diesen Aufgaben die Werte in Tabelle 18 aufgrund einer eingeschränkten Datenbasis entstanden sind und somit als nicht bedrohende Einschränkungen betrachtet werden können. Bei der Bohnenaufgabe müsste die Stichprobengrösse wesentlich erhöht werden ($N \geq 190$), damit der Zusammenhang signifikant werden würde, da hier der vorzufindende Effekt gering ist (vgl. Tab. 18). Da bei der Bohnenaufgabe jedoch nach Ausschluss von Ausreissern ein ausschliesslich positives Konfidenzintervall beobachtet werden konnte, können auch bei dieser Aufgabe die Werte in Tabelle 18 als nicht bedrohende Einschränkung betrachtet werden. Auch bei der Pulveraufgabe müsste die Stichprobengrösse wesentlich erhöht werden ($N \geq 550$), damit der Zusammenhang signifikant werden würde, da hier der Effekt sehr gering ist (vgl. Tab. 18). Da bei der Pulveraufgabe jedoch auch nach Ausschluss von Ausreissern das Konfidenzintervall nicht ausschliesslich positiv ist, liegen somit bei dieser Aufgabe bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) bedrohende Einschränkungen vor.

Tabelle 19: Plausibilität der Konzepte im Bereich Messwiederholung (MW) und Mengenvergrösserung (GM) und Rückmeldungen der Expertinnen und Experten in den offenen Antwortfenstern.

	Intendierte Konzepte	Mittelwert und Standardabweichung bzgl. der Plausibilität		Rückmeldungen der Expertinnen und Experten in den offenen Antwortfenstern
Ahorn	MW	***M* = 3.5**	*SD* = 0.8	**A.** Der Beschleunigungsweg des Samens muss berücksichtigt werden. Das Messen mit doppelter Pulthöhe und das Halbieren der Zeit ist daher eigentlich nicht korrekt, da somit davon ausgegangen wird, dass die Geschwindigkeit beim Fallen annähernd gleich bleibt. *(Rückmeldung einer Expertin und eines Experten)* **B.** Die Ungenauigkeit der Stoppuhr wird von der Ungenauigkeit der Handmessung überdeckt. Somit ist es zwar korrekt, wenn bei der Wahl des Messinstruments auf die Genauigkeit von Stoppuhr B verwiesen wird, im Kontext ist dies aber eigentlich irrelevant. *(Rückmeldung eines Experten)*
	GM	*M* = 2.1	*SD* = 1.1	
Bohne	MW	*M* = 2.1	*SD* = 0.8	**C**. Das Durchführen von Messwiederholungen ist hier eigentlich nicht zentral, da die Einzelmessung bereits sehr ungenau ist. Das Messen mit einer grossen Menge ist hingegen zwingend, da nur so eine Messung gemacht werden kann. *(Rückmeldung eines Experten)*
	GM	***M* = 3.9**	*SD* = 0.4	
Faden	MW	***M* = 3.5**	*SD* = 0.8	**D.** Die Lernenden sind sich wahrscheinlich nicht bewusst, dass das Messen mit einem Faden oder doppeltem Faden[40] unterschiedliche Ergebnisse ergibt. Sie entscheiden sich vermutlich eher unbewusst für eine der beiden Varianten. *(Rückmeldung von zwei Experten)*
	GM	*M* = 1.6	*SD* = 0.7	

[40] Mit einem Faden messen: Faden an der Federwaage anknoten und daran ziehen.

Mit doppeltem Faden messen: Faden in Form einer Schlaufe über die Federwaage legen und an beiden Enden ziehen.

Filzstift	MW	*M* = 2.4	*SD* = 0.9	**E.** Die Lernenden sind sich nicht bewusst, dass der Effekt (Wandern von Filzstiftpunkten auf Papier) vom Löschpapier abhängt und die Ergebnisse somit ungenau sind. Darum werden sie vermutlich keine Messwiederholungen machen, v.a. auch weil eine Messung gemäss Aufgabe ziemlich lange dauert (zwei Minuten). Darum werden sie wahrscheinlich eher mit mehreren Punkten messen. *(Rückmeldung eines Experten)* **F.** Wenn sich die Lernenden Gedanken darüber machen, ob die gemessene Zeit und die zurückgelegte Strecke sich linear verhalten, werden sie eher nicht mit mehr als zwei Minuten messen. *(Rückmeldung einer Expertin)*
	GM *(mit mehreren Punkten messen)*	***M* = 3.1**	*SD* = 0.6	
	GM *(mit mehr als zwei Minuten messen)*	*M* = 2.1	*SD* = 0.8	
Münze	MW	*M* = 2.0	*SD* = 0.8	**G**. Das Durchführen von Messwiederholungen ist hier eigentlich nicht zentral, da die Einzelmessung bereits sehr ungenau ist. Das Messen mit einer grossen Menge ist hingegen zwingend, da nur so eine Messung gemacht werden kann. *(Rückmeldung eines Experten)*
	GM	***M* = 3.8**	*SD* = 0.5	
Pulver	MW	***M* = 3.1**	*SD* = 0.6	**H.** Das Phänomen (Veränderung der Temperatur des Wassers beim Auflösen von Pulver) ist den Lernenden vermutlich nicht bekannt und darum werden sie eher nicht mit mehreren Tütchen Pulver aufs Mal messen, da sie nicht wissen, dass die Menge an Pulver einen Einfluss haben könnte (fehlendes Fachwissen). *(Rückmeldung eines Experten)*
	GM	*M* = 2.5	*SD* = 1.1	

Anmerkung: Fett hervorgehoben sind Mittelwerte von mindestens 3.0.

Als eine Bedingung, dass bei den Aufgaben von Hinweisen für kognitive Validität aus dem Bereich (III) ausgegangen wird, musste im Rahmen vorliegender Arbeit mindestens eines der beiden Konzepte (MW oder GM) von den Expertinnen und Experten als eher naheliegend (Mittelwert ≥ 3.0) eingeschätzt werden (Bedingung 1). Es muss bei den Aufgaben jeweils nur eines der Konzepte als eher nahliegend eingeschätzt werden, da zur Erreichung von Qualitätsstandard 3 (vgl. Unterkapitel 2.3.2) nur eines der Konzepte vorausgesetzt wurde, wobei grundsätzlich immer beide durchführbar sind. Als weitere Bedingung für Hinweise für kognitive Qualität aus dem Bereich (III) mussten die Rückmeldungen der Expertinnen und Experten zu den Aufgaben und intendierten Konzepten als *nicht* bedrohende Einschränkungen beurteilt werden (Bedingung 2). Beispiele für bedrohende Einschränkungen wären, wenn im Expertenrating rückgemeldet wird, dass eine Aufgabe unverständlich ist oder Fremdwörter enthält oder wenn die Konzepte im Bereich Messwiederholung oder Messen mit einer Menge *beide* bei einer Aufgabe als nicht naheliegend oder prinzipiell nicht durchführbar erachtet werden.

Beim Betrachten von Tabelle 19 kann festgestellt werden, dass bei einigen Aufgaben des Problemtyps «Messen» (Ahorn, Faden, Pulver) eher die Konzepte im Bereich Messwiederholung als naheliegend beurteilt wurden (Mittelwert ≥ 3.0), während bei den anderen Aufgaben (Bohne, Filzstift und Münze) eher die Konzepte im Bereich Mengenvergrösserung als plausibel eingeschätzt wurden. In der Folge werden die Ergebnisse für diese beiden Gruppen von Aufgaben vorgestellt.

Intendierte Konzepte im Bereich Messwiederholung (MW) eher naheliegend:
Bei der Ahorn-, Faden- und Pulveraufgabe wurden die intendierten Konzepte im Bereich Messwiederholung von den Expertinnen und Experten als eher naheliegend eingeschätzt (Ahorn, $M = 3.5$; Faden, $M = 3.5$; Pulver, $M = 3.1$): Somit scheint es gemäss Einschätzungen der Expertinnen und Experten plausibel zu sein, dass die Lernenden bei diesen Aufgaben Messungen wiederholen, um ein möglichst genaues Ergebnis zu erhalten. Die intendierten Konzepte im Bereich Mengenvergrösserung (GM) wurden bei diesen Aufgaben als eher nicht naheliegend beurteilt (Ahorn, $M = 2.1$; Faden, $M = 1.6$; Pulver, $M = 2.5$): Somit scheint es gemäss Einschätzungen der Expertinnen und Experten eher nicht naheliegend zu sein, dass die Schülerinnen und Schüler bei diesen Aufgaben mit einer Menge messen, um ein möglichst genaues Ergebnis zu erhalten. Dass die Konzepte im Bereich Mengenvergrösserung von den Expertinnen und Experten als eher nicht naheliegend beurteilt wurden, stellt jedoch keine bedrohende Einschränkung im Validitätsargumentationsprozess dar, da für die Suche nach Hinweisen für

kognitive Validität aus dem Bereich (III) nur eines der beiden Konzepte (MW oder GM) als eher nahliegend eingeschätzt werden musste (vgl. Bedingung 1).

Des Weiteren wurden die Rückmeldungen der Expertinnen und Experten bei der Ahorn-, Faden- und Pulveraufgabe (Rückmeldung A, B, D und H, vgl. Tab. 19) als nicht bedrohende Einschränkungen beurteilt (vgl. Bedingung 2). Rückmeldung A wurde als nicht bedrohende Einschränkung eingestuft, da bei den Aufgaben des Problemtyps «Messen» die methodischen und nicht die fachinhaltlichen Kompetenzen im Vordergrund stehen (vgl. Unterkapitel 6.4.1) und es somit in Ordnung ist, wenn angenommen wird, dass die Geschwindigkeit beim Fallen des Ahornsamens annähernd gleich bleibt. Auch Rückmeldung B wurde als nicht bedrohende Einschränkung interpretiert. Zwar wird die Ungenauigkeit der Stoppuhr bei der Ahornaufgabe durch die Ungenauigkeit der Handmessung überdeckt, dennoch ist die Wahl des genaueren Messinstruments auch bei dieser Aufgabe zentral, um zusätzliche Ungenauigkeiten aufgrund des Messinstruments zu vermeiden (vgl. auch intendiertes Konzept «Für eine genaue Messung sollte immer ein möglichst genaues Messinstrument gewählt werden», Unterkapitel 6.4.1). Auch Rückmeldung D wurde als nicht bedrohende Einschränkung eingestuft: Dass die Schülerinnen und Schüler unbewusst mit einem Faden oder doppeltem Faden messen, wurde bei der Beurteilung der Kompetenzen berücksichtigt (z. B. bei der Vorgehensweise oder beim Messergebnis (Toleranzbereich wurde für beide Varianten definiert); vgl. auch kodierte Kriterien in Unterkapitel 2.3.2) und zudem wurden die Lernenden durch Aufträge im vorstrukturierten Schülerprotokoll («Mit wie vielen Fäden hast du gemessen?»; vgl. Unterkapitel 6.4.2) darauf hingewiesen, dass mit unterschiedlich vielen Fäden aufs Mal gemessen werden kann. Rückmeldung H wurde auch als nicht bedrohende Einschränkung interpretiert, denn diese Rückmeldung zeigt lediglich auf, dass es gemäss Einschätzung des Experten eher nicht naheliegend ist, dass die Schülerinnen und Schüler mit mehreren Tütchen Pulver aufs Mal messen, da ihnen hierzu das Fachwissen fehlt. Da für Bedingung 1 jedoch nur eines der beiden Konzepte (MW oder GM) als plausibel eingeschätzt werden musste und bei der Pulveraufgabe das Durchführen von Messwiederholungen als eher naheliegend eingestuft wurde, stellt dies keine bedrohende Einschränkung dar.

Intendierte Konzepte im Bereich Mengenvergrösserung (GM) eher naheliegend: Bei der Bohnen-, Filzstift- und Münzenaufgabe wurden die intendierten Konzepte im Bereich Mengenvergrösserung als eher naheliegend beurteilt (Bohne, $M = 3.9$; Filzstift (mit mehreren Punkten messen), $M = 3.1$[41]; Münze, $M = 3.8$): Somit scheint es gemäss Einschätzungen der Expertinnen und Experten plausibel zu sein, dass die Lernenden bei diesen Aufgaben mit einer Menge messen, um ein möglichst genaues Ergebnis zu erhalten. Die intendierten Konzepte im Bereich Messwiederholung wurden bei der Bohnen-, Filzstift- und Münzenaufgabe als eher nicht naheliegend eingeschätzt (Bohne, $M = 2.1$; Filzstift, $M = 2.4$; Münze, $M = 2.0$): Somit scheint es gemäss Einschätzungen der Expertinnen und Experten eher nicht naheliegend zu sein, dass die Lernenden bei diesen Aufgaben Messwiederholungen durchführen, um ein möglichst genaues Ergebnis zu erhalten. Da zur Erfüllung von Bedingung 1 aber nur eines der beiden Konzepte (MW oder GM) als eher naheliegend beurteilt werden musste, stellt dies keine bedrohende Einschränkung im Validitätsargumentationsprozess dar.

Zudem wurden die Rückmeldungen der Expertinnen und Experten bei der Bohnen-, Filzstift- und Münzenaufgabe (Rückmeldung C, E, F und G, vgl. Tab. 19) als nicht bedrohende Einschränkungen eingeschätzt (vgl. Bedingung 2). Rückmeldung C, E, und G wurden jeweils als nicht bedrohende Einschränkungen beurteilt, da diese Rückmeldungen lediglich aufzeigen, dass das Durchführen von Messwiederholungen bei diesen Aufgaben als eher nicht naheliegend eingeschätzt wurde, da zum Beispiel die Einzelmessung bereits sehr ungenau ist (Bohnen- und Münzenaufgabe) oder eine Messung viel Zeit beansprucht (Pulveraufgabe). Da gemäss Bedingung 1 aber nur eines der beiden Konzepte (MW oder GM) als eher naheliegend beurteilt werden musste und bei diesen Aufgaben jeweils das Messen mit einer Menge als plausibel eingeschätzt wurde, stellt dies keine bedrohende Einschränkung dar. Auch Rückmeldung F wurde als nicht bedrohende Einschränkung betrachtet. Grundsätzlich sind bei der Filzstiftaufgabe beide Strategien der Mengenvergrösserung (‚mit mehreren Punkten messen' oder ‚mit mehr als zwei Minuten messen') möglich und wurden deshalb bei der Kodierung der Daten und im Expertenrating berücksichtigt. Tatsächlich scheint aber die Strategie der Mengenvergrösserung durch das Messen mit mehr als zwei Minuten wenig naheliegend zu sein, während das Messen mit mehreren Punkten

[41] Bei der Filzstiftaufgabe gibt es zwei Möglichkeiten der Mengenvergrösserung: Es kann mit mehreren Punkten oder mit mehr als zwei Minuten gemessen werden. Das Messen mit mehreren Punkten wurde als eher naheliegend beurteilt ($M = 3.1$), während das Messen mit mehr als zwei Minuten als eher nicht naheliegend ($M = 2.1$) beurteilt wurde.

plausibel erscheint[42]. Da bei der Filzstiftaufgabe aber nur eine der beiden Möglichkeiten im Bereich Mengenvergrösserung als naheliegend beurteilt werden musste (vgl. Bedingung 1) und das Messen mit mehreren Punkten als plausibel eingeschätzt wurde, stellt Rückmeldung F ebenfalls keine bedrohende Einschränkung dar.

Insgesamt können somit bei allen Aufgaben des Problemtyps «Messen» Hinweise für kognitive Validität aus dem Bereich (III) gefunden werden. Bei allen Aufgaben konnte gezeigt werden, dass eines der beiden Konzepte (MW oder GM) als eher naheliegend (Mittelwert ≥ 3) eingeschätzt wurde und somit Bedingung 1 erfüllt ist. Zudem wurden alle Rückmeldungen der Expertinnen und Experten als nicht bedrohende Einschränkungen eingeschätzt und somit ist auch Bedingung 2 erfüllt. Infolgedessen gibt es bei den Aufgaben des Problemtyps «Messen» Hinweise für kognitive Validität aus dem Bereich (III), da einerseits eine Aktivierung der intendierten Konzepte im Bereich Messwiederholung oder Mengenvergrösserung grundsätzlich als plausibel erscheint und zudem von den Expertinnen und Experten keine Aspekte erkannt werden konnten, welche die kognitive Validität beeinträchtigen könnten.

8.5 Fazit und Diskussion

Teilstudie II untersucht FF2, indem analysiert wird, inwiefern Hinweise für kognitive Validität aus den Bereichen (I) bis (III) bei den Aufgaben des Problemtyps «Messen» gefunden werden können (vgl. Kapitel 5). In der Folge werden zuerst Hinweise erläutert, die auf eine kognitiv valide Testwertinterpretation hindeuten. Daraufhin werden bedrohende Einschränkungen aufgeführt. Schlussendlich findet eine zusammenfassende Beurteilung im Hinblick darauf statt, inwiefern die Aufgaben anscheinend kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen.

Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (I) konnte festgestellt werden, dass bei den Aufgaben des Problemtyps «Messen» beim Lösen der Aufgaben in vielen Bereichen offenbar mehrheitlich über ein intendiertes Konzept nachgedacht wurde und somit die intendierten Konzepte seitens der Lernenden aktiviert zu werden scheinen (bei 13 von 18 untersuchten Bereichen, vgl. Tab. 17). In manchen Bereichen konnte anhand der kategorisierten Schüleraussagen nicht festgestellt werden, über welche Konzepte die Lernenden beim Lösen

[42] Dies zeigt sich auch in den Ergebnissen: Während 13 Schülerinnen und Schüler mit mehreren Punkten aufs Mal messen, denkt anscheinend nur ein Schüler über das Messen mit mehr als zwei Minuten nach.

der Aufgabe nachgedacht haben (bei 4 von 18 untersuchten Bereichen). Dies stellt jedoch keine bedrohende Einschränkung bei der Validitätsargumenation dar. Zudem konnte gezeigt werden, dass beim Problemtyp «Messen» höhere Denkprozesse (hoher Q.i.K.-Score) mit einer besseren Lösung der Aufgabe einhergehen (höheres Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle; $r = .46$, $p \leq .001$) und somit anscheinend das Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle die intendierten Konzepte abbilden kann (Hinweis für kognitive Validität aus dem Bereich (II)). Dass auf Ebene der einzelnen Aufgaben dieser positive Zusammenhang vorerst nicht immer beobachtet werden konnte (Bohnen-, Faden-, Filzstift-, Münzen- und Pulveraufgabe), konnte in den meisten Fällen (Bohnen-, Faden-, Filzstift-, Münzenaufgabe) auf die eingeschränkte Datenbasis auf Aufgabenebene zurückgeführt werden und stellt somit keine bedrohende Einschränkung dar. Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (III) konnte festgestellt werden, dass bei den Aufgaben jeweils das intendierte Konzept im Bereich Messwiederholung (Ahorn-, Faden- und Pulveraufgabe) oder Mengenvergrösserung (Bohnen-, Filzstift- und Münzenaufgabe) als eher naheliegend eingeschätzt wurde und es somit plausibel scheint, dass diese Konzepte grundsätzlich seitens der Lernenden aktiviert werden können. Zudem wurden im Rating keine Aspekte erkannt, welche die kognitive Validität beeinträchtigen könnten.

Bei der Pulveraufgabe konnten bedrohende Einschränkungen aufgedeckt werden. Einerseits konnte festgestellt werden, dass beim Lösen der Aufgabe die Mehrheit der Schülerinnen und Schüler im Bereich der Wahl des Messinstruments anscheinend über kein intendiertes Konzept nachgedacht hat, sondern andere Aspekte handlungsleitend waren (Thermometer B wurde oft nicht aufgrund der Messgenauigkeit gewählt, sondern aus praktischen Gründen). Zudem konnte bei der Pulveraufgabe, auch nach Ausschluss von Ausreissern, kein ausschliesslich positives Konfidenzintervall für r beobachtet werden. Dies bedeutet, dass anscheinend qualitativ höhere Denkprozesse bei den intendierten Konzepten nicht zwingend mit einer besseren Lösung der Aufgabe einhergehen und somit das Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle die intendierten Konzepte nicht zufriedenstellend abbilden kann.

Zusammenfassend kann somit davon ausgegangen werden, dass die Ahorn-, Bohnen-, Faden-, Filzstift- und Münzenaufgabe kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen, da hier einige Hinweise für kognitive Validität aus den Bereichen (I) bis (III) aufgedeckt werden konnten und keine bedrohenden Einschränkungen vorliegen. Somit scheinen sich diese Aufgaben zur

Diagnose experimenteller Kompetenzen im Bereich des naturwissenschaftlichen Messens von Lernenden des 8. Schuljahres zu eignen. Zudem konnte im Rahmen von Teilstudie II festgestellt werden, dass bei der Ahorn- und Fadenaufgabe vor allem das Durchführen von Messwiederholungen plausibel ist (Ahorn, $M = 3.5$; Faden, $M = 3.5$) und sich diese Aufgaben folglich insbesondere zur Diagnose experimenteller Kompetenzen in diesem Bereich eignen. Andererseits scheinen sich die Bohnen-, Filzstift- und Münzenaufgabe insbesondere für die Diagnose experimenteller Kompetenzen im Bereich der Strategie Mengenvergrösserung zu eignen (Plausibilität: Bohne, $M = 3.9$; Filzstift, $M = 3.1$; Münze, $M = 3.8$). Die Pulveraufgabe hingegen scheint keine kognitiv validen Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden zuzulassen, da hier im Rahmen des Validierungsprozesses bedrohende Einschränkungen aufgedeckt werden konnten. Die Pulveraufgabe müsste somit, falls sie erneut zur Diagnose experimenteller Kompetenzen eingesetzt werden sollte, angepasst werden, vor allem im Bereich der Wahl des Messinstruments.

Neben der Frage, inwiefern Hinweise für kognitive Validität aus den Bereichen (I) bis (III) bei den Aufgaben des Problemtyps «Messen» gefunden werden können, bestand ein weiteres Forschungsziel vorliegender Arbeit darin, ein analytisches Verfahren zu entwickeln, mit welchem Hinweise für kognitive Validität aus den Bereichen (I) bis (III) aufgedeckt werden können (vgl. Kapitel 5). Teilstudie II zeigt somit auch, dass sich das entwickelte Verfahren eignet, um Hinweise für kognitive Validität aus den Bereichen (I) bis (III) aufzudecken. Durch das Expertenrating konnte geprüft werden, inwiefern die kategorisierten Schüleraussagen den intendierten Konzepten entsprechen, inwiefern die intendierten Konzepte bei den Aufgaben naheliegend sind und ob bei den Aufgaben Aspekte erkannt werden können, welche die kognitive Validität beeinträchtigen könnten. Dadurch konnten Hinweise für kognitive Validität aus dem Bereich (I) und (III) generiert werden. Anhand der im Expertenrating eingeschätzten Qualität bei den intendierten Konzepten konnte zudem untersucht werden, inwiefern qualitativ höhere Denkprozesse mit einer besseren Lösung der Aufgabe einhergehen. Dadurch konnten Hinweise für kognitive Validität aus dem Bereich (II) generiert werden. Insgesamt konnten so durch das angewendete analytische Verfahren Hinweise für kognitive Validität aus den Bereichen (I) bis (III) aufgedeckt werden und folglich scheint sich das Verfahren für die Suche nach Indizien für kognitive Validität aus den Bereichen (I) bis (III) zu eignen.

9. Zusammenfassung und Ausblick

Naturwissenschaftliche Bildung ist von zentraler Bedeutung, um am gesellschaftlichen Leben teilzuhaben. Dabei umfasst naturwissenschaftliche Bildung nicht nur die Kenntnis inhaltsbezogener Kompetenzen, sondern auch naturwissenschaftliche Denk- und Arbeitsweisen (vgl. z. B. D-EDK, 2016; NRC, 2012; KMK, 2005) und hier insbesondere experimentelle Kompetenzen. Bei der Frage, was experimentelle Kompetenzen sind und welche Teilkompetenzen diese umfassen (vgl. z. B. Höttecke & Rieß, 2015; Metzger et al., 2019), wurde in vorliegender Arbeit von einem eher breiten Begriffsverständnis von Experimentieren ausgegangen (vgl. Unterkapitel 2.1), eine handelnde Auseinandersetzung allerdings immer vorausgesetzt. Somit wurden auch Vorgehensweisen wie eine naturwissenschaftliche Messung oder ein kriteriengeleiteter Vergleich zu den experimentellen Zugängen gezählt und experimentelle Teilkompetenzen wurden als die Fähigkeit aufgefasst, unterschiedliche experimentelle Problemstellungen (z. B. eine Messung oder einen Vergleich) lösen zu können (vgl. z. B. auch Gut, Metzger, et al., 2014). Im Rahmen dieser als Problemtypenansatz (vgl. Unterkapitel 2.2) bezeichneten Modellierung experimenteller Kompetenzen wurde bei vorliegender Arbeit auf den Problemtypen «Messen» fokussiert. Im Zentrum stand dabei die Diagnose experimenteller Kompetenzen, wofür Tests mit Realexperimenten als Benchmark betrachtet (vgl. z. B. Wenning, 2007; Baxter & Shavelson, 1994) und verschiedene Erhebungsmethoden eingesetzt werden können (vgl. Unterkapitel 3.2). Entsprechend wurden in Teilstudie I vorliegender Arbeit verschiedene Erhebungsmethoden bei Tests mit Realexperimenten bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose verglichen. Zudem stellt sich bei der Diagnose stets auch die Frage, inwiefern die Ergebnisse der Diagnose valide Schlüsse bezüglich der Kompetenzen der Lernenden im untersuchten Bereich zulassen, wobei für eine umfassende Validitätsbeurteilung verschiedene Validitätsaspekte berücksichtigt werden können (vgl. Messick 1995 und Kapitel 4). Im Rahmen vorliegender Arbeit wurde auf die kognitive Validität fokussiert und deshalb in Teilstudie II ein Testverfahren zur Diagnose experimenteller Kompetenzen bei Tests mit Realexperimenten am Beispiel der Aufgaben des Problemtyps «Messen» kognitiv validiert. Damit leistet vorliegende Arbeit einen zentralen Beitrag zu einer umfassenden Validitätsbeurteilung im Rahmen der Gesamtvalidierungsstudie des Projekts ExKoNawi (vgl. Unterkapitel 6.1). In der Folge werden die Ergebnisse der Teilstudien I und II zusammengefasst sowie Implikationen für den naturwissenschaftlichen Unterricht und die fachdidaktische Forschung hergeleitet.

In Teilstudie I wurden die Ergebnisse der Kompetenzdiagnose bei Tests mit Realexperimenten anhand verschiedener Erhebungsmethoden verglichen. Hierzu wurden Schülerprotokolle (P), Videoaufnahmen während des Experimentierens (V) und Interviews (I) untersucht. Während Videoaufnahmen und Interviews zeitintensive Methoden darstellen, bieten Schülerprotokolle eine ökonomische Alternative. Bei Schülerprotokollen stellt sich jedoch die Frage, wie sichergestellt werden kann, dass die Schülerinnen und Schüler protokollieren, was sie tatsächlich gemacht haben (vgl. z. B. Gut-Glanzmann, 2012). Konkret wurde im Rahmen von Teilstudie I der Frage nachgegangen (vgl. Kapitel 5, FF1), inwiefern zusätzliche Erhebungsmethoden zum Schülerprotokoll (PV, PI) die Genauigkeit der Diagnostik erhöhen und inwiefern die untersuchten Erhebungsmethoden (P, PV und PI) bezüglich der Genauigkeit des Ergebnisses an den gesetzten Benchmark (PVI) herankommen. Die Ergebnisse von Teilstudie I zeigen, dass die experimentellen Kompetenzen durch zusätzliche Interviews genauer diagnostiziert werden als ohne Interviews. Dieses Ergebnis ist von zentraler Bedeutung für den naturwissenschaftlichen Unterricht, da es aufzeigt, dass es trotz des organisatorischen und zeitlichen Mehraufwands notwendig ist, die Lernenden manchmal zu ihren experimentellen Handlungen zu befragen. So kann mehr über experimentelle Handlungen, die womöglich nicht ausreichend in den Schülerprotokollen festgehalten wurden, und über die Gedanken von Lernenden in Erfahrung gebracht werden. Dieser Befund bestätigt frühere Studien, die darauf hinweisen, dass Schülerprotokolle nur einen eingeschränkten Einblick in die experimentellen Kompetenzen von Schülerinnen und Schülern zulassen (vgl. z. B. Abrahams et al., 2013; Gott & Dugan, 2002; Gut-Glanzmann, 2012) und dass zwischen schriftlichen Materialien und den Gedanken von Lernenden nicht stets ein systematischer Zusammenhang besteht (vgl. z. B. Vorholzer et al., 2020). Entsprechend kann davon ausgegangen werden, dass sich diese Erkenntnis auch auf andere Testinstrumente mit Realexperimenten übertragen lässt, sodass auch bei diesen durch zusätzliche Interviews ein genaueres Ergebnis der Kompetenzdiagnose erzielt werden könnte. Qualitativ wurde im Rahmen von Teilstudie I untersucht, inwiefern die genauere Diagnose experimenteller Kompetenzen anhand zusätzlicher Interviews augenscheinlich mit anderen Variablen begründet werden kann. Die qualitative Betrachtung zeigte, dass der Messzeitpunkt einen Einfluss auf das genauere Ergebnis der Diagnose anhand zusätzlicher Interviews zu haben scheint, sodass bei den späteren Schulbesuchen die Ergebnisse der Diagnose mit und ohne zusätzliche Interviews besser übereinzustimmen scheinen. Dies könnte auf einen Gewöhnungseffekt an die Protokollmethode hindeuten, wodurch die Schülerinnen und Schüler nach einer ausreichenden Gewöhnung die Schülerprotokolle anscheinend genauer führen und somit die Ergebnisse

der Kompetenzdiagnose mit und ohne zusätzliche Interviews besser übereinzustimmen scheinen. Gestützt wird diese Folgerung zusätzlich dadurch, dass auch Emden und Sumfleth (2012) darauf hinweisen, dass vor allem bei leistungsschwächeren Schülerinnen und Schülern eine sorgfältige Einführung in die Protokollmethode zentral ist. Die qualitative Betrachtung zeigte zudem, dass vor allem auch bei leistungsschwächeren Schülerinnen und Schülern und / oder Lernenden mit geringer Motivation Interviews zu einem genaueren (bzw. höheren) Ergebnis der Diagnose führen. Inwiefern hier vor allem die kognitiven oder sprachlichen Fähigkeiten eine Rolle spielen oder auch motivationale Faktoren entscheidend sind, muss systematisch untersucht werden, zum Beispiel mit Hilfe einer mehrfaktoriellen Varianzanalyse. Eine solche Erkenntnis kann Indizien darüber liefern, bei welchen Gruppen von Schülerinnen und Schülern (z. B. kognitiv leistungsstärkere Schülerinnen und Schüler oder eher motivierte Lernende) sich Schülerprotokolle womöglich noch *eher* für eine genaue Diagnose experimenteller Kompetenzen eignen. Die Ergebnisse von Teilstudie I haben zudem gezeigt, dass zusätzliche Videoaufnahmen von den Schülerinnen und Schülern während des Experimentierens keinen entscheidenden Vorteil bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose zu bringen scheinen, insbesondere dann nicht, wenn zusätzliche Interviews durchgeführt wurden. Somit kommt PI bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose an den gesetzten Benchmark (PVI) heran. Dieses Ergebnis deckt sich mit Befunden von anderen Studien: Einige Studien konnten zeigen, dass bei Tests mit Realexperimenten die Diagnose experimenteller Kompetenzen durch Beobachtungen von Schülerinnen und Schülern während des Experimentierens und durch Schülerprotokolle zu ähnlichen Ergebnissen führen (vgl. z. B. Baxter et al., 1992; Emden & Sumfleth, 2012; Shavelson et al., 1991, 1993). Dieser Befund ist von zentraler Bedeutung für zukünftige Studien und den naturwissenschaftlichen Unterricht, da er eine Entlastung bedeuten kann: Das Ergebnis deutet darauf hin, dass es nicht zwingend notwendig scheint, dass die Schülerinnen und Schüler vermehrt während des Experimentierens beobachtet werden. Dennoch sollte nicht vollständig auf die Beobachtung verzichtet werden, da durch die Beobachtung mehr über handlungsbezogene experimentelle Kompetenzen, wie zum Beispiel die Handhabung von Messinstrumenten, oder mögliche Schwierigkeiten der Lernenden beim Experimentieren in Erfahrung gebracht werden kann.

In Teilstudie II wurde ein Testverfahren zur Diagnose experimenteller Kompetenzen bei Tests mit Realexperimenten am Beispiel der Aufgaben des Problemtyps «Messen» kognitiv validiert. Validität ist abhängig vom sozialen Kontext und dem Verwendungszweck (vgl. z. B. Kane, 2006; Messick, 1995) und

kann somit nicht isoliert betrachtet werden, sondern muss im Zusammenhang mit der Anwendung untersucht werden. Im Rahmen vorliegender Arbeit wurde untersucht, inwiefern die Aufgaben des Problemtyps «Messen» kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen von Lernenden des 8. Schuljahres im Bereich des naturwissenschaftlichen Messens im Sinne einer Standortbestimmung zulassen. Die in der Folge aufgeführten Erkenntnisse sind somit vor diesem Hintergrund zu verstehen. Im Rahmen des Validierungsprozesses wurde nach Hinweisen für kognitive Validität aus drei Bereichen gesucht: (I) Die Schülerinnen und Schüler denken beim Lösen der Aufgaben anscheinend mehrheitlich über ein intendiertes Konzept nach, (II) höhere Denkprozesse gehen mit einer besseren Lösung der Aufgabe einher und (III) die intendierten Konzepte werden als naheliegend eingeschätzt (Plausibilität) und es werden keine Aspekte erkannt, welche die kognitive Validität beeinträchtigen könnten. Bei der Suche nach Hinweisen für kognitive Validität aus den Bereichen (I) bis (III) konnte festgestellt werden, dass anscheinend die Ahorn-, Bohnen-, Faden-, Filzstift- und Münzenaufgabe kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulassen. Einerseits konnte gezeigt werden, dass bei diesen Aufgaben in vielen Bereichen die Mehrheit der Schülerinnen und Schüler beim Lösen der Aufgabe anscheinend über ein intendiertes Konzept nachgedacht hat (Hinweis aus dem Bereich (I)) oder es wurde nicht ersichtlich, worüber die Lernenden nachdenken, was jedoch keine bedrohende Einschränkung bei der Validitätsargumentation darstellt. Bei der Suche nach Hinweisen für kognitive Validität aus dem Bereich (II) konnte zudem festgestellt werden, dass auf der Ebene des Problemtyps «Messen» ($N = 108$) und bei der Ahornaufgabe ($N = 18$) qualitativ höhere Denkprozesse (hoher Q.i.K.-Score) mit einem höheren Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle einhergehen und somit zwischen diesen zwei Massen ein positiver Zusammenhang besteht. Bei der Bohnen-, Faden-, Filzstift- und Münzenaufgabe konnte festgestellt werden, dass die nicht ausschliesslich positiven Korrelationen (Konfidenzintervall nicht nur positiv) wahrscheinlich aufgrund der eingeschränkten Datenbasis auf Aufgabenebene ($N = 18$) entstehen und somit als nicht bedrohende Einschränkungen gedeutet werden können, welche die Validitätsargumentation nicht bedeutsam schwächen. Zudem haben die Ergebnisse von Teilstudie II auch gezeigt, dass bei den Aufgaben des Problemtyps «Messen» jeweils das intendierte Konzept im Bereich Messwiederholung (MW; Ahorn, Faden, Pulver) oder Mengenvergrösserung (GM; Bohnen, Filzstift, Münze) von den Expertinnen und Experten als eher naheliegend eingeschätzt wurde und es somit grundsätzlich plausibel scheint, dass diese Konzepte seitens der Lernenden aktiviert werden können (Hinweis aus dem Bereich (III)).

Da zur Erreichung von Qualitätsstandard 3 nur eines der beiden Konzepte (MW oder GM) vorausgesetzt wurde (vgl. Unterkapitel 2.3.2), stellt es keine bedrohende Einschränkung dar, dass bei den Aufgaben jeweils nur eines der Konzepte als eher naheliegend beurteilt wurde. Die aufgedeckten Indizien für kognitive Validität aus den Bereichen (I) bis (III) deuten insgesamt darauf hin, dass die Ahorn-, Bohnen-, Faden-, Filzstift- und Münzenaufgaben kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen von Lernenden im Bereich des naturwissenschaftlichen Messens zulassen. Dieser Befund ist von besonderer Relevanz für den naturwissenschaftlichen Unterricht und die fachdidaktische Forschung, da somit eine Möglichkeit zur Diagnose experimenteller Kompetenzen in Tests mit Realexperimenten aufgezeigt werden kann, die kognitiv valide Schlüsse bezüglich der Kompetenzen von Lernenden des 8. Schuljahres im Bereich des naturwissenschaftlichen Messens zulässt. Bei der Pulveraufgabe hingegen konnten im Rahmen des Validierungsprozesses bedrohende Einschränkungen aufgedeckt werden: Somit scheint die Pulveraufgabe keine kognitiv validen Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zuzulassen. Einerseits hat Teilstudie II gezeigt, dass die Mehrheit der Schülerinnen und Schüler im Bereich der Wahl des Messinstruments anscheinend beim Lösen der Aufgabe über kein intendiertes Konzept nachgedacht hat, sondern andere Aspekte handlungsleitend waren. Zudem konnte bei der Pulveraufgabe, auch nach Ausschluss von Ausreissern, kein ausschliesslich positiver Zusammenhang zwischen dem Q.i.K.-Score und dem Ergebnis der Kompetenzdiagnose anhand der Schülerprotokolle beobachtet werden und somit scheinen höhere Denkprozesse bei den intendierten Konzepten nicht zwingend mit einer besseren Lösung der Aufgabe einherzugehen. Infolgedessen scheinen bei der Pulveraufgabe die Schülerprotokolle die intendierten Konzepte nicht ausreichend abzubilden. Da somit Indizien vorliegen, die aufzeigen, dass die Pulveraufgabe womöglich keine kognitiv validen Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden zulässt, müsste die Aufgabe für einen zukünftigen Einsatz angepasst werden, vor allem im Bereich der Wahl des Messinstruments. Hierfür könnten zum Beispiel zwei Thermometer zur Verfügung gestellt werden, die beide ungefähr die gleiche Länge aufweisen und sich somit nicht in ihrer Praktikabilität (Becher mit Wasser kippt beim Messen um) unterscheiden (vgl. Unterkapitel 6.4.1, Abb. 4).

Da im Rahmen von Teilstudie II gezeigt wurde, dass die Pulveraufgabe anscheinend keine kognitiv validen Schlüsse bezüglich der experimentellen Kompetenzen der Lernenden im Bereich des naturwissenschaftlichen Messens zulässt, wurden die Befunde von Teilstudie I erneut unter Ausschluss der Pulveraufgabe

geprüft ($N = 90$). Dabei konnte festgestellt werden, dass auch unter Ausschluss der Pulveraufgabe zusätzliche Interviews zu einem genaueren Ergebnis der Kompetenzdiagnose führen: Auf Ebene der Stichprobe konnten bedeutsame Mittelwertsunterschiede zwischen P und PI beobachtet werden ($p \leq .001$; $d = 0.9$) und auf Individualebene konnte keine hinreichend hohe Korrelation zwischen P und PI festgestellt werden ($r = .62$; $p \leq .001$). Zudem scheinen Videoaufnahmen, auch unter Ausschluss der Pulveraufgabe, keinen entscheidenden Vorteil bezüglich der Genauigkeit des Ergebnisses der Kompetenzdiagnose zu bringen, insbesondere dann nicht, wenn zusätzliche Interviews durchgeführt wurden. Somit unterscheiden sich auf Ebene der Stichprobe die Mittelwerte von PI und PVI nicht signifikant ($p = .32$) und auf Individualebene konnte eine hohe Korrelation zwischen PI und PVI beobachtet werden ($r = .99$; $p \leq .001$), was für eine Austauschbarkeit der Methoden spricht. Die Befunde von Teilstudie I können somit auch unter Ausschluss der Pulveraufgabe bestätigt werden.

Ein weiteres Ziel von Teilstudie II bestand darin, ein analytisches Verfahren zum Generieren von Hinweisen für kognitive Validität aus den Bereichen (I) bis (III) zu entwickeln und auf seine Eignung zu prüfen, denn in vielen Studien werden im Rahmen des Validierungsprozesses nur Hinweise für kognitive Validität aus dem Bereich (I) berücksichtigt (z. B. Dickmann, Eickhorst, Theyβen, et al., 2014; Hadenfeldt et al., 2014; Kröger, 2019). Im Zuge der Untersuchungen von Teilstudie II konnte festgestellt werden, dass sich das Verfahren zum Generieren von Hinweisen für kognitive Validität aus den Bereichen (I) bis (III) eignet, da bei allen Aufgaben Hinweise aufgedeckt werden konnten, die (1) für eine kognitiv valide Testwertinterpretation sprechen, (2) als nicht bedrohende Einschränkungen bei der Validitätsargumentation gedeutet werden können und / oder (3) bedrohende Einschränkungen darstellen. Somit leistet Teilstudie II einen Forschungsertrag über die untersuchten Aufgaben hinaus, da ein analytisches Verfahren zum Generieren von Hinweisen für kognitive Validität aus den Bereichen (I) bis (III) aufgezeigt werden konnte.

Im Rahmen von Teilstudie I wurden die Videoaufnahmen vor allem produktorientiert ausgewertet, der Fokus lag also auf den Ergebnissen (z. B. korrekte Experimentdurchführung). Um mehr über die experimentellen Handlungen der Lernenden und mögliche Schwierigkeiten beim Experimentieren zu erfahren, können die erhobenen Daten im Sinne eines fortführenden Validierungsprozesses zusätzlich prozessorientiert ausgewertet werden. Da sich dafür insbesondere Videoaufnahmen von Schülerinnen und Schülern während des Experimentierens eignen, sollen die Videoaufnahmen in einem nächsten Schritt zusammen mit den Schülerprotokollen (PV) prozessorientiert ausgewertet werden. Die

Berücksichtigung von Videoaufnahmen *und* Schülerprotokollen ist wichtig, da die Lernenden die Aufgaben in Einzelarbeit bearbeitet haben und dabei nicht zum Lauten Denken aufgefordert wurden und somit wichtige Hinweise, zum Beispiel im Bereich der Experimentplanung, nicht im Video ersichtlich werden. Bei der prozessorientierten Auswertung können zum Beispiel Lernprozessgrafiken von den Aufgaben des Problemtyps «Messen» erstellt werden (vgl. auch z. B. Emden & Sumfleth, 2012; Klos et al., 2008; Walpuski, 2006), wobei experimentelle Handlungen, wie beispielsweise die Experimentplanung oder Experimentdurchführung, im zeitlichen Verlauf visuell dargestellt werden. Ergänzend kann durch Farbcodes auch die Richtigkeit der experimentellen Handlungen visualisiert werden. So kann untersucht werden, ob die Lernenden bei den Aufgaben des Problemtyps «Messen» zielführend (vor der Experimentdurchführung wird in der Lernprozessgrafik eine Experimentplanung visualisiert) oder eher ausprobierend vorgehen, inwiefern ihre Vorgehensweisen logisch schlüssig sind (z. B. nach einer korrekten Experimentplanung wird das Experiment auch richtig durchgeführt) und in welchen Bereichen sie allenfalls Schwierigkeiten haben (nicht korrekte Handlungen in der Lernprozessgrafik). Diese Erkenntnisse können als wichtige zusätzliche Indizien im Rahmen eines fortführenden Validierungsprozesses dienen, da sie aufzeigen, wie die Lernenden die Aufgaben des Problemtyps «Messen» bearbeitet haben (z. B. entsprechen die real ablaufenden Bearbeitungsprozesse den intendierten Prozessen) und wo sie allenfalls Schwierigkeiten haben.

Im Rahmen von Teilstudie I und II wurden Erhebungsmethoden verglichen und ein Testverfahren am Beispiel der Aufgaben des Problemtyps «Messen» kognitiv validiert. Im Zuge der Auswertungen wurden jedoch nicht mögliche Kompetenzen und Konzepte seitens der Schülerinnen und Schüler im Bereich des naturwissenschaftlichen Messens detailliert analysiert und beschrieben. Dies soll in einem nächsten Schritt erfolgen. Es gibt wenige Arbeiten, die systematisch mögliche Kompetenzen und Konzepte im Bereich des naturwissenschaftlichen Messens von Schülerinnen und Schülern der Sekundarstufe I untersucht haben (vgl. Hellwig, 2012), sondern häufig wird auf die Zielgruppe von Studierenden fokussiert. Beispielhafte Studien, welche Kompetenzen von *Schülerinnen und Schülern* untersucht haben, sind die von Kanari und Millar (2004), Lubben und Millar (1996) oder Munier und anderen (2013). Hellwig (2012) hat zudem ein Sachstrukturmodell im Bereich Messunsicherheiten unter anderem für die Sekundarstufe I entwickelt. Dieses kann als Ausgangslage zum Entwickeln von Lernumgebungen, zum Erheben von Konzepten seitens der Lernenden oder zur Formulierung von Kompetenzen dienen. Schulz (2021) hat aufbauend auf dem Modell von

Hellwig (2012) Kompetenzen im Bereich Messunsicherheiten für Studienanfänger beschrieben und hierzu ein Testinstrument entwickelt. Die im Rahmen von Teilstudie I und II vorliegender Arbeit erhobenen Daten können somit eine Anschlussmöglichkeit an die aktuelle Forschung bieten (vgl. z. B. Hellwig, 2012 und Schulz, 2021), indem anhand dieser Daten ein erster Überblick über mögliche Kompetenzen und Konzepte von Schülerinnen und Schülern der Sekundarstufe I im Bereich des naturwissenschaftlichen Messens gewonnen werden kann. Anhand der Interviews und der Einschätzungen im Expertenrating kann beispielsweise untersucht werden, in welchen Bereichen die Lernenden bereits eher naturwissenschaftlich adäquate Vorstellungen haben (intendierte Konzepte von eher hohem Niveau) und in welchen Bereichen die Konzepte der Schülerinnen und Schüler womöglich noch nicht naturwissenschaftlich adäquaten Vorstellungen entsprechen (intendierte Konzepte von eher niedrigem Niveau oder falsche Bezüge zum Konzept). Ergänzend können auch die allgemeineren Fragen aus den Interviews (vgl. Unterkapitel 6.4.4 und Anhang, Teil A)[43] und die Auswertungen hierzu genutzt werden. Durch diese Fragen kann noch mehr über die Konzepte der Lernenden erfahren werden. Anhand der Videoaufnahmen von den Schülerinnen und Schülern während des Lösens der Aufgaben des Problemtyps «Messen» kann zudem mehr über handlungsbezogene experimentelle Kompetenzen im Bereich des naturwissenschaftlichen Messens in Erfahrung gebracht werden. Hierfür können die Videos zum Beispiel im Hinblick darauf analysiert werden, wie die Lernenden mit den Messinstrumenten umgehen oder wie die Lernenden Messwiederholungen durchführen (z. B. ob Messwiederholungen annähernd auf die gleiche Weise durchgeführt werden). Zudem können die Auswertungen im Bereich von Qualitätsstandard 5 (vgl. Unterkapitel 2.3.2) genutzt werden, um zu untersuchen, welche Quellen für Messunsicherheiten die Schülerinnen und Schüler hauptsächlich benennen und welche Lösungsvorschläge zur Steigerung der Messgenauigkeit sie daraus ableiten. Die so gewonnenen Erkenntnisse können als ein erster Überblick über mögliche Konzepte und Kompetenzen von Schülerinnen und Schülern der Sekundarstufe I im Bereich des naturwissenschaftlichen Messens und somit als Ausgangslage für eine umfassendere Erforschung dieser dienen.

[43] Z. B. waren im Interviewleitfaden auch allgemeinere Fragen zum Umgang mit Messreihen und Ausreissern integriert, in Anlehnung an Hellwig (2012) oder Lubben und Millar (1996).

10. Literaturverzeichnis

Abrahams, I., Reiss, M., & Sharpe, R. (2013). The assessment of practical work in school science. *Studies in Science Education*, *49*(2), 209–251.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421.

AREA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arndt, K. (2016). *Experimentierkompetenz erfassen—Analyse von Prozessen und Mustern am Beispiel von Lehramtstudierenden der Chemie*. Berlin: Logos Verlag.

Arnold, J., Kremer, K., & Mayer, J. (2014). Understanding Students' Experiments—What kind of support do they need in inquiry tasks? *International Journal of Science Education*, *36*(16), 2719–2749.

Barzel, B., Reinhoffer, B., & Schrenk, M. (2012). Experimentieren im mathematisch-naturwissenschaftlichen Unterricht—Schüler lernen wissenschaftlich zu denken und arbeiten. In W. Rieß, M. Wirtz, B. Barzel, & A. Schulz (Hrsg.), *Das Experiment im Unterricht* (S. 103–127). Münster: Waxmann.

Baxter, G. P., Elder, A. D., & Glaser, R. (1995). *Cognitive Analysis of Science Performance Assessment* (CSE Technical Report Nr. 398). National Center for Research on Evaluation, Standards, and Student Testing, Universitiy of California.

Baxter, G. P., & Glaser, R. (1998). Investigating the Cognitive Complexity of Science Assessment. *Educational Measurement: Issues and Practice*, *17*(3), 37–45.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, *21*(3), 279–298.

Baxter, G. P., Shavelson, R. J., Goldmann, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, *29*, 1–17.

Bonetti, A., Gut, C. & Metzger S. (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). Universität Regensburg.

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). Universität Regensburg.

Bonetti, A. In Vorbereitung. *Experimentieren als Diagnoseinstrument - Validierung eines hands-on Experimentiertests für die Sekundarstufe 1 mit Aufgabenkontexten aus der Chemie, Physik und Biologie*. Dissertation.

Bortz, J., & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung* (3. Aufl.). Berlin, Heidelberg: Springer Medizin Verlag.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin, Heidelberg: Springer.

Coelho, S. M., & Séré, M.-G. (1998). Pupils' Reasoning and Practice during Hands-on Activities in the Measurement Phase. *Research in Science and Technological Education*, *16*(1), 79–90.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, N.J.: L. Erlbaum Associates.

Deardorff, D. L. (2001). *Introductory Physics Students' Treatment of Measurement Uncertainty* [Degree of Doctor of Philosophy, North Carolina State University]. Zugriff am 10.02.2022. Verfügbar unter: http://isobudgets.com/pdf/papers/10_DeardorffDissertation.pdf

Deutschschweizer Erziehungsdirektorenkonferenz (D-EDK). (2016). *Lehrplan 21—Fachbereich NMG*. Zugriff am 10.02.2022. Verfügbar unter: http://v-ef.lehrplan.ch/container/V_EF_DE_Fachbereich_NMG.pdf

Dickmann, M. (2016). *Messung von Experimentierfähigkeiten—Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*. Berlin: Logos Verlag.

Dickmann, M., Eickhorst, B., Theyßen, H., Neumann, K., Schecker, H., & Schreiber, N. (2014). Measuring experimental skills in large-scale assessments: Developing a simulation-based test instrument. In C. P. Constantinou, N. Papadouris, & A. Hadjigeorgiou (Hrsg.), *E-Book proceedings of the ESERA 2013 conference: Science education research for evidence-based teaching and coherence in learning: Bd. Part 11.* (S. 50-58). Nicosia, Cyprus: ESERA.

Duit, R., Gropengiesser, H., & Stäudel, L. (2007). *Naturwissenschaftliches Arbeiten. Unterricht und Material 5-10.* Seelze-Velber: Erhard Friedrich Verlag.

Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens*. Berlin: Logos Verlag.

Emden, M., & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung—Zur Eignung einer Protokollmethode für die Bewertung von Experimentierprozessen. *MNU*, *65*(2), 68–75.

Erb, M., & Bolte, C. (2011). Kompetenzdiagnostik im Bereich Naturwissenschaftliche Erkenntnisgewinnung. In D. Höttecke (Hrsg.), *Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie*. Münster: Lit Verlag.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251.

Erziehungsdirektorenkonferenz (EDK) (Hrsg.). (2011). *Grundkompetenzen für die Naturwissenschaften - Nationale Bildungsstandards.* Zugriff am: 01.03.2022. Verfügbar unter: Grundkompetenzen für die Naturwissenschaften: nationale Bildungsstandards. Frei gegeben von der EDK-Plenarversammlung am 16.6.2011 (edudoc.ch)

Fairbrother, R., & Hackling, M. (1997). Is this the right answer? *International Journal of Science Education*, *19*(8), 887–894.

Funke, J., & Spering, M. (2006). Methoden der Denk- und Problemlöseforschung. In J. Funke (Hrsg.), *Denken und Problemlösen* (1. Aufl., S. 647–744). Göttingen, Bern, Toronto, Seattle: Hogrefe.

Garden, R. (1999). Development of TIMSS Performance Assessment Tasks. *Studies in Educational Evaluation*, *25*(3), 217–241.

Gott, R., & Dugan, S. (1995). *Investigative work in the science curriculum*. Buckingham: Open University Press.

Gott, R., & Dugan, S. (2002). Problems with the Assessment of Performance in Practical Science: Which way now? *Cambridge Journal of Education*, *32*(2), 183–201.

Gott, R., Dugan, S., Roberts, R., & Hussain, A. (2003). *Research into understanding scientific evidence*. Zugriff am 10.02.2022. Verfügbar unter: http://community.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Gott, R., & Duggan, S. (2002). Problems with the Assessment of Performance in Practical Science: Which way now? *Cambridge Journal of Education*, *32*(2), 183–201.

Grube, C. (2011). *Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung. Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I.* Dissertation, Universität Kassel.

Gut, C., Hild, P., Metzger, S., & Tardent, J. (2014). Projekt ExKoNawi: Modell für hands-on Assessments experimenteller Kompetenz. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht.* (S. 171–173). Kiel: IPN.

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. 328–331. Universität Regensburg.

Gut, C., Labudde, P., & Ramseier, E. (2010). Large-Scale Experimentiertests: Ansätze zur Analyse von Intemschwierigkeiten. In D. Höttecke (Hrsg.), *Entwicklung naturwissenschaftlichen Denkens zwischen Phänomen und Systematik* (S. 245–247). Münster: Lit Verlag.

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

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

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

Haag, G., Scheid, J., Löffler, P., & Kauertz, A. (2018). Desiderata bei der manuellen Ausführung von Experimenten. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht—Normative und empirische Dimensionen.* (S. 847–850). Universität Regensburg.

Hadenfeldt, J. C., Repenning, B., & Neumann, K. (2014). Die kognitive Validität von Ordered Multiple Choice Aufgaben zur Erfassung des Verständnisses von Materie. *Zeitschrift für Didaktik der Naturwissenschaften*, *20*(1), 57–68.

Häder, M. (2010). Erhebungsmethoden. In *Empirische Sozialforschung—Eine Einführung* (2. Aufl., S. 187–338). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hafner, T. (2012). *Proportionalität und Prozentrechnung in der Sekundarstufe I - empirische Untersuchung und didaktische Analysen* (G. Kaiser, R. Borromeo Ferri, & W. Blum, Hrsg.). Wiesbaden: Vieweg und Teubner Verlag / Springer Fachmedien.

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

Hammann, M., Phan, T. T. H., Ehmer, M., & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42*(2), 66–72.

Heinicke, S. (2012). *Aus Fehlern wird man klug—Eine Genetisch-Didaktische Rekonstruktion des «Messfehlers»*. Berlin: Logos Verlag.

Helfferich, C. (2011). *Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews.* (4. Aufl.). Wiesbaden: Verlag für Sozialwissenschaften.

Heller, K. A., & Perleth, C. (2000). *KFT 4-12+ R. Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz Test.

Hellwig, J. (2012). *Messunsicherheiten verstehen: Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtfaches Physik* [Erlangung des Grades «Doktor der Naturwissenschaften»]. Ruhr-Universität Bochum.

Henke, Christian. (2007). *Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. Untersuchung am Beispiel des HIGHSEA Projekts in Bremerhaven.* Berlin: Logos Verlag.

Hild, P., Brückmann, M., & Gut, C. (2017). Aussagen zur Konstruktvalidität beim experimentellen Problemtyp „Effektbasiertes Vergleichen". In C. Maurer (Hrsg.): *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis.* (S. 332–335). Universität Regensburg.

Höttecke, D., & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung—Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*, 127–139.

Kampach, M. (2018). *Experimentierprozesse von Lehramtsstudierenden der Biologie—Eine Videostudie*. Berlin: Logos Verlag.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. Brennan (Hrsg.), *Educational Measurement* (4. Aufl., S. 17–64). Westport CT: American Council on Education and Praeger.

Kirchner, S., & Priemer, B. (2010). Welche Kompetenzen zeigen Schüler beim Umgang mit Variablen? In D. Höttecke (Hrsg.), *Entwicklung naturwissenschaftlichen Denkens zwischen Phänomen und Systematik*. Münster: Lit Verlag.

Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, Massachusetts, London: MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science*, *12*, 1–48.

Klos, S. (2008). *Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht – der Einfluss eines integrierten Unterrichtskonzepts.* Berlin: Logos Verlag.

Klos, S., Henke, C., Kieren, C., Walpuski, M., & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen—Zwei verschiedene Kompetenzen. *Zeitschrift für Pädagogik*, *54*(3), 304–321.

Koenen, J. (2014). *Entwicklung und Evaluation von experimentunterstützenden Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen*. Berlin: Logos Verlag.

Konrad, K. (2020). Lautes Denken. In *Handbuch Qualitative Forschung in der Psychologie. Band 2: Design und Verfahren* (S. 373–393). Wiesbaden: Springer.

Kröger, J. (2019). *Struktur und Entwicklung des Professionswissens angehender Physiklehrkräfte*. (Dissertation), Kiel: Christian-Albrechts-Universität.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Leuders, T. (2014). Modellierung mathematischer Kompetenzen—Kritierien für eine Validitätsprüfung aus fachdidaktischer Sicht. *Journal für Mathematik-Didaktik*, *35*(1), 7–48.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*(8), 955–968.

Maiseyenka, V. (2014). *Modellbasiertes Experimentieren im Unterricht*. Berlin: Logos Verlag.

Mannel, S. (2011). *Assessing scientific inquiry – Development and evaluation of a test for the low-performing stage*. Berlin: Logos Verlag.

Masnick, A. M., & Morris, B. J. (2002). Reasoning from data: The effect of sample size and variability on children's and adults' conclusions. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 643–648.

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.

Mayering, H., & Wimmer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2-9*. Bern: Hogrefe.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Metzger, S., & Gut, C. (2017). Symposium: Experimentelle Kompetenzen in den Naturwissenschaften (ExKoNawi). In C. Maurer (Hrsg.): *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis.* (324–327). Universität Regensburg.

Metzger, S., Gut, C., Hild, P., & Tardent, J. (2014). Modelling and assessing experimental competence: An interdisciplinary progress model for hands-on assessments. *E-Proceedings of the ESERA 2013 Conference*.

Metzger, S., Lembens, A., & Arnold, J. (2019). Praktisches naturwissenschaftliches Arbeiten im Spannungsfeld der Disziplinen. In S. Habig (Hrsg.): *Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen.* (S. 60–65). Universität Duisburg-Essen.

Millar, R., & Osborn, J. (1998). *Beyond 2000: Science education for the future*. King's College London, School of Education.

Miller, D. M., & Linn, R. L. (2000). Validation of Performance-Based Assessments. *Psychological Measurement*, *24*(4), 367–378.

Munier, V., Merle, H., & Brehelin, D. (2013). Teaching scientific measurement and uncertainty in elementary school. *International Journal of Science Education*, *35*(16), 2752–2783.

National Research Council (NRC). (2012). *A framework for K-12 science education. Practices, crosscutting concepts, and core ideas.* Washington D.C.: National Academies Press.

Nehring, A., Nowak, K. H., Upmeier zu Belzen, A., & Tiemann, R. (2014). Ausgewählte Analysen der «VerE-Studie»—Zur Trennbarkeit und zu Zusammenhängen von Fachwissen und Kompetenzen im Bereich Erkenntnisgewinnung. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht.* (S. 177–179). Kiel: IPN.

Neumann, K. (2004). *Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker*. Berlin: Logos Verlag

Reynolds, C. R., Livingston, R., & Willson, V. (2010). *Measurement and assessment in education* (2. Aufl.). London: Pearson Education International.

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, M. A. (2000). *On the «Exchangeability» of Hands-On and Computer-Simulated Science performance Assessments* (Nr. 531; CSE Technical Report). National Center for Research on Evaluation, Standards, and Student Testing.

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), 1045–1063.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the Validity of Cognitive Interpretations of Scores From Alternative Concept-Mapping Techniques. *Educational Assessment*, *7*(2), 99–141.

Schaper, N. (2009). Aufgabenfelder und Perspektiven bei der Kompetenzmodellierung und -messung in der Lehrerbildung. *Lehrerbildung auf dem Prüfstand*, *2*(1), 166–199.

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

Schiepe-Tiska, A., Rönnebeck, S., Schöps, K., Neumann, K., Schmidtner, S., Parchmann, I., & Prenzel, M. (2016). Naturwissenschaftliche Kompetenz in PISA 2015 – Ergebnisse des internationalen Vergleichs mit einem modifizierten Testansatz. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Hrsg.), *PISA 2015—Eine Studie zwischen Kontinuität und Innovation* (S. 45–98). Münster: Waxmann.

Schreiber, N. (2012). *Diagnostik experimenteller Kompetenz—Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*. Berlin: Logos Verlag.

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.

Schulz, A., Wirtz, M., & Starauscheck, E. (2012). Das Experiment in den Naturwissenschaften. In W. Rieß, M. Wirtz, B. Barzel, & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht: Schüler lernen wissenschaftlich denken und arbeiten* (S. 15–38). Münster, New York, München, Berlin: Waxmann.

Schulz, J. (2021). *Entwicklung eines Testinstrumentes zur Erfassung von Kompetenzen im Umgang mit Messunsicherheiten*. Dissertation, Mathematisch-Naturwissenschaftlichen Fakultät der Humboldt-Universität.

Schwippert, K., Kasper, D., Köller, O., McElvany, N., Selter, C., Steffensky, M., & Wendt, H. (Hrsg.). (2020). *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster, New York: Waxmann.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2005). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss (Jahrgangsstufe 10)*. München: Luchterhand.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215–232.

Shavelson, R. J., Baxter, G. P., Pine, J., Yuré, J., Goldmann, S. R., & Smith, B. (1991). Alternative Technologies for Large Scale Science Assessment: Instrument of Education Reform. *School Effectiveness and School Improvement*, *2*(2), 97–114.

Shavelson, R. J. & Ruiz-Primo, M. A. (1999). Leistungsbewertung im naturwissenschaftlichen Unterricht. *Unterrichtswissenschaft, 27*(2), 102-127.

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.

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

Stebler, R., Reusser, K., & Ramseier, E. (1998). Praktische Anwendungsaufgaben zur integrierten Förderung formaler und materialer Kompetenzen—Erträge aus dem TIMSS-Experimentiertest. *Bildungsforschung und Bildungspraxis*, *20*(1), 28–54.

Stocké, V. (2019). Persönlich-mündliche Befragung. In N. Bauer & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (2. Aufl., S. 745–756). Wiesbaden: Springer Fachmedien.

Suida, R., & Grabowski, A. (2012). Combined measurements – A way to improve the measurement accuracy of an additive quantity. *Measurement*, *45*, 1165–1169.

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

Varelas, M. (1997). Third and Fourth Graders' Conceptions of Repeated Trials and Best Representatives in Science Experiments. *Journal of Research in Science Teaching*, 853–872.

Völzke, K. (2012). Lautes Denken bei kompetenzorientierten Diagnoseaufgaben zur naturwissenschaftlichen Erkenntnisgewinnung. *Reihe Studium und Forschung,* (20), Universität Kassel: Zentrum für Lehrerbildung und Forschung, 1–106.

von Aufschnaiter, C., & Rogge, C. (2010). Wie lassen sich Verläufe der Entwicklung von Kompetenz modellieren? *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 95-114.

Vorholzer, A., Hägele, J., & von Aufschnaiter, C. (2020). Instruktionen kohärent anlegen und Kompetenzaufbau untersuchen: Zugänge und Herausforderungen am Beispiel experimentbezogener Kompetenz. *Unterrichtswissenschaft*, *48*(1), 61–89.

Vorholzer, A., & von Aufschnaiter, C. (2020). Dimensionen und Ausprägungen fachinhaltlicher Kompetenz in den Naturwissenschaften – ein Systematisierungsversuch. *Zeitschrift für Didaktik der Naturwissenschaften*, *26*(1), 1–18.

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*, *22*, 25–41.

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

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The Dependability and Interchangeability of Assessment Methods in Science. *Applied Measurement in Education*, *3*(3), 277–301.

Wellnitz, N., & Mayer, J. (2012). Beobachten, Vergleichen und Experimentieren: Wege der Erkenntnisgewinnung. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik 5. Didaktik der Biologie—Standortbestimmung und Perspektiven* (S. 63–80). Innsbruck: Studien Verlag.

Wellnitz, N., & Mayer, J. (2013). Erkenntnismethoden in der Biologie—Entwicklung und Evaluation eines Kompetenzmodells. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 315–345.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientif literacy. *Journal of Physics Education Online*, *4*(2), 21–24.

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

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen.* Göttingen: Hogrefe, Verlag für Psychologie.

Wolming, S., & Wikstrom, C. (2010). The Concept of Validity in Theory and Practice. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 17–132.

11. Anhang

Übersicht

Teil A Interviewleitfaden zu den Aufgaben des Problemtyps «Messen» am Beispiel der Fadenaufgabe

Teil B Kategoriensystem zur Auswertung der Interviews im Rahmen von Teilstudie II am Beispiel der Fadenaufgabe

Teil C Expertenrating zur Einschätzung der kategorisierten Schüleraussagen

A Interviewleitfaden zu den Aufgaben des Problemtyps «Messen» am Beispiel der Fadenaufgabe

Zu jeder Aufgabe des Problemtyps «Messen» wurde ein Interviewleitfaden entwickelt. Die Interviewleitfäden sind für die verschiedenen Aufgaben des Problemtyps «Messen» jeweils identisch aufgebaut. Der Aufbau wird in der Folge am Beispiel der Fadenaufgabe illustriert.
Im Interviewleitfaden sind einerseits Fragen integriert, anhand welcher die Schülerinnen und Schüler aufgefordert werden, ihre Vorgehensweisen und Überlegungen beim Lösen der Aufgaben des Problemtyps «Messen» zu verbalisieren. Bei diesen Fragen wurde während des Interviews jeweils Bezug zu den von den Schülerinnen und Schülern ausgefüllten Schülerprotokollen genommen. Zudem waren im Interviewleitfaden einige allgemeinere Fragen integriert, um mögliche Konzepte seitens der Schülerinnen und Schüler im Bereich des naturwissenschaftlichen Messens zu erheben. Diese allgemeineren Fragen wurden in Anlehnung an die Arbeiten von Hellwig (2012) sowie Lubben und Millar (1996) entwickelt. Bei den allgemeineren Fragen wurden den Schülerinnen und Schülern oft laminierte Karten vorgelegt, zum Beispiel mit einer vorgegebenen Messreihe, anhand welcher sie die Vorgehensweise zur Ermittlung eines Schlussresultats erklären sollen.

Anmerkungen:

SoS	steht für *Schülerin oder Schüler*
SuS	steht für *Schülerinnen und Schüler*
i07 etc.	diese Verweise stehen für die entsprechenden Stellen / Indikatoren im vorstrukturierten Schülerprotokoll

Leitfaden Interview Faden

Name SoS	
Code	
Schule, Klasse	
Datum, Uhrzeit	

Einleitung für die SuS:
Du hast ein Experiment durchgeführt, um herauszufinden, bei welcher Belastung ein Faden reisst. Wir werden jetzt zu diesem Experiment gemeinsam ein Interview durchführen. Das Interview wird auf Video aufgezeichnet. Wichtig ist, dass ich dein Video nicht weitergeben oder weiterzeigen werde. Mir geht es darum, im Interview herauszufinden, was du dir beim Experiment überlegt hast und wie du vorgegangen bist. Dazu werde ich dir einige Fragen stellen.

Durchführungshinweise für den Interviewer / die Interviewerin:
Mit den Schülerinnen und Schülern wird während des Interviews gemeinsam ihr ausgefülltes Schülerprotokoll durchgegangen, d.h. dass während des Interviews die entsprechenden Seiten im Schülerprotokoll aufgeschlagen werden. Zu den jeweiligen Aufträgen im Schülerprotokoll werden im Interview gezielt Fragen gestellt, bei denen die Schülerinnen und Schüler ihre Handlungen erklären und ihre Gedanken verbalisieren sollen.

QS 1 (adäquate Vorgehensweise) und QS 2 (eindeutiges Ergebnis mit korrekter Einheit)

Wenn im Schülerprotokoll bei i01 / i02 eine Vorgehensweise beschrieben wurde.	**Wenn im Schülerprotokoll bei i01 / i02 keine Vorgehensweise beschrieben wurde.**
Du hast hier deine Vorgehensweise aufgeschrieben. Warum bist du genau so vorgegangen?	Erkläre mir, wie du vorgegangen bist und warum du genau so vorgegangen bist?

Wenn im Schülerprotokoll ein Resultat aufgeschrieben wurde (i01 / i02).	**Wenn im Schülerprotokoll kein Resultat aufgeschrieben wurde (i01 / i02).**
Wie bist du auf dieses Resultat gekommen?	Hast du ein Resultat erhalten? **Wenn ja**: Welches? Wie bist du auf dieses Resultat gekommen? **Wenn nein**: Warum hast du kein Resultat erhalten?

Wenn im Schülerprotokoll ein Resultat *mit* Einheit aufgeschrieben wurde (i01 / i02).	**Wenn im Schülerprotokoll keine Einheit aufgeschrieben wurde (i01 / i02).**
Du hast die Einheit _____aufgeschrieben. Wieso hast du diese Einheit genommen?	Du hast aufgeschrieben / gesagt, dass dein Resultat _____ist. Welche Einheit hat dieses Resultat? Warum hat es diese Einheit?

Allgemeine Fragen	
	Zeige mir, wie du den Faden für deine Messungen an der Federwaage befestigt hast. **(1 Faden-Ansatz / 2 Faden-Ansatz)** **(Federwaage A & B, Faden und Schere zur Verfügung stellen)**
	Max, Pascal und Cornelia haben das gleiche Experiment gemacht und folgende Belastungen erhalten: Max: **1500 g** Pascal: **1300 g** Cornelia: **1400 g** Wer hat richtig gemessen? Warum? **(Den SuS die laminierte Karte «Resultate» zeigen)**

QS 3a (Messstrategie: Messwiederholung)

Allgemeine Fragen (Einstieg)	
	Ist es bei dieser Aufgabe besser einmal oder mehrmals zu messen? Warum?
	Warum wird bei manchen Experimenten mehr als einmal gemessen?

Wenn im Schülerprotokoll bei i07 eine Antwort aufgeschrieben wurde.	**Wenn im Schülerprotokoll bei i07 keine Antwort aufgeschrieben wurde.**
Du schreibst hier, dass du _____mal die Belastung gemessen hast. Warum hast du genau _____mal gemessen? **Wenn mehrmals gemessen wurde: (i09)** Du hast gesagt, dass du _____ mal die Belastung gemessen hast. Wie sind deine Ergebnisse von den verschiedenen Messungen? Erkläre, wie du dann auf ein Endresultat gekommen bist.	Wie viele Male hast du gemessen, bis du dein Endresultat hattest? Warum hast du genau _____mal gemessen? **Wenn mehrmals gemessen wurde: (i09)** Du hast gesagt, dass du _____ mal die Belastung gemessen hast. Wie sind deine Ergebnisse von den verschiedenen Messungen? Erkläre, wie du dann auf ein Endresultat gekommen bist.

Allgemeine Fragen

Julian hat dreimal hintereinander gemessen. Die Belastungen betrugen dabei:

Messung Nr.	Belastungen
1	**1500 g**
2	**1400 g**
3	**1500 g**

Wie kann er mit diesen Resultaten zu einem Endresultat kommen?

(Den SuS die laminierte Karte «Messwiederholungen, 1.» zeigen und einen Taschenrechner zur Verfügung stellen)

Wenn man noch mehr Messungen macht, zum Beispiel zehn Messungen anstelle von drei, wird dann das Ergebnis genauer oder bleibt es gleich genau? Warum?

Iva hat beim Experiment folgende Ergebnisse erhalten:

Messung Nr.	Belastungen
1	**1300 g**
2	**1400 g**
3	**1300 g**
4	**2500 g**
5	**1500 g**

Wie kommt sie zu einem Endresultat?

(Den SuS die laminierte Karte «Messwiederholungen, 2.» zeigen und einen Taschenrechner zur Verfügung stellen)

Fabian und Alexandra haben beide viermal gemessen. Sie haben folgende Belastungen erhalten:

Fabian	
Messung Nr.	Belastung
1	**1400 g**
2	**1500 g**
3	**1300 g**
4	**1400 g**

Alexandra	
Messung Nr.	Belastung
1	**1800 g**
2	**1400 g**
3	**1100 g**
4	**1300 g**

Welchen Resultaten kannst du mehr vertrauen? Warum?

(Den SuS die laminierte Karte «Messwiederholungen, 3.» zeigen)

Frage zur Variablenkontrolle
Wenn man mehrmals misst, muss man dann beides Mal die gleiche Federwaage nehmen oder kann man bei jeder Messung eine andere Federwaage nehmen? Warum?

QS 3b (Messstrategie: Messen mit einer (grossen) Menge)

Allgemeine Fragen (Einstieg)	
	Spielt es bei dieser Aufgabe eine Rolle, ob man den Faden wie in Bild 1 oder wie in Bild 2 befestigt? Warum? Bild 1  Bild 2 **(Den SuS die laminierte Karte «Menge, 1.» zeigen)**
	Werden sich die Ergebnisse unterscheiden, je nachdem ob man wie in Bild 1 oder wie in Bild 2 misst? **(Den SuS die laminierte Karte «Menge, 1.» zeigen)** **Wenn ja, sie werden sich unterscheiden:** - Warum werden sie sich unterscheiden? - Wie werden sie sich unterscheiden?
	Stell dir vor, Paul hat wie in Bild 2 gemessen und dabei eine Belastung von 2700 g erhalten. Was ist das Endergebnis? **(Den SuS die laminierte Karte «Menge, 2.» zeigen)**
	Ist es bei dieser Aufgabe besser so (wie in Bild 1) oder so (wie in Bild 2) zu messen? Warum? **(Den SuS die laminierte Karte «Menge, 1.» zeigen)**

Wenn im Schülerprotokoll bei i08 eine Antwort vorhanden ist	**Wenn im Schülerprotokoll bei i08 nicht steht, mit wie vielen Fäden aufs Mal gemessen wurde**
Du schreibst, dass du mit einem Faden / _____Fäden aufs Mal gemessen hast. Warum hast du genau mit einem Faden / ____ Fäden aufs Mal gemessen? **Wenn mit mehreren Fäden gemessen wurde: (i09)** Welches Ergebnis hast du dann für ____ Fäden erhalten? Erkläre, wie du dann vorgegangen bist, um die Belastung, bei welcher ein Faden reisst, zu erhalten.	Mit wie vielen Fäden hast du aufs Mal gemessen? Warum hast du genau einen Faden / _____ Fäden aufs Mal genommen? **Wenn mit mehreren Fäden gemessen wurde: (i09)** Welches Ergebnis hast du dann für ____ Fäden erhalten? Erkläre, wie du dann vorgegangen bist, um die Belastung, bei welcher ein Faden reisst, zu erhalten.

QS 4 (Wahl Messinstrument)

Fragen zum Messinstrument (i05 / i06)	
	Hast du mit Federwaage A oder mit Federwaage B gemessen? Warum hast du dich für diese Federwaage entschieden? **Wenn ein SoS sagt, dass er beide Federwaagen verwendet hat:** Warum hast du beide Federwaagen genommen?
	Können beide Federwaagen gleich genau messen oder messen sie unterschiedlich genau? Woran erkennst du das? **(Federwaagen zur Verfügung stellen)**

QS 5 (Lösungsvorschläge zur Steigerung der Messgenauigkeit)

Fragen zum Schülerprotokoll (i03)	
	Ist dein Resultat genau oder ungenau? Warum? **Wenn nur wenig Gründe für die Genauigkeit / Ungenauigkeit genannt werden:** Kannst du dir weitere Gründe vorstellen, warum dein Ergebnis genau / ungenau ist?

Allgemeine Frage	
	Ava hat als Endresultat eine Belastung von 1433.33 g aufgeschrieben. Ist es sinnvoll so viele Stellen nach dem Dezimalpunkt anzugeben oder würden auch weniger Stellen nach dem Punkt ausreichen? Warum? **(Den SuS die laminierte Karte «Anzahl Stellen» zeigen)**

Fragen zum Schülerprotokoll (i04)	
	Kannst du Vorschläge machen, wie man noch genauer messen könnte? **Wenn wenige Vorschläge genannt werden:** Gibt es weitere Möglichkeiten, wie man noch genauer messen könnte?

Allgemeine Fragen zum Abschluss des Interviews	
	Kommt es für die Genauigkeit darauf an, ob man Federwaage A oder B wählt? Warum? **Wenn genannt wird, dass Federwaage A genauer ist:** Kann das Resultat, auch wenn mit Federwaage A gemessen wurde, dennoch ungenau sein? Warum?
	Wenn man mehrmals misst, wird dann bei dieser Aufgabe das Resultat genauer? Warum?
	Wenn man mit mehreren Fäden aufs Mal misst, wird dann das Resultat genauer? Warum? **(Vorzeigen, was «Messen mit mehreren Fäden» bedeutet (z. B. Messen mit Schlaufe vorzeigen))**

B Kategoriensystem zur Auswertung der Interviews im Rahmen von Teilstudie II am Beispiel der Fadenaufgabe

Im Rahmen von Teilstudie II wurden die Interviews mit Hilfe eines Kategoriensystems ausgewertet. Das Kategoriensystem ist für alle Aufgaben des Problemtyps «Messen» identisch aufgebaut: Der Aufbau wird in der Folge am Beispiel der Fadenaufgabe illustriert.

Das Kategoriensystem beinhaltet Oberkategorien im Bereich der bepunkteten Qualitätsstandards (QS) des Problemtyps «Messen», also Oberkategorien zu QS 1 bis 4 (vgl. Unterkapitel 2.3.2). Für die Auswertung von FF2 (vgl. Kapitel 5) wurden die Oberkategorien im Bereich von QS 3 und QS 4 berücksichtigt, da hier die intendierten Konzepte der Aufgaben des Problemtyps «Messen» liegen. Somit wurden in den Bereichen Messwiederholung und Mengenvergrösserung jeweils zwei Oberkategorien und im Bereich Wahl des Messinstruments eine Oberkategorie für die Auswertungen von FF2 berücksichtigt. Diese Oberkategorien sind in der folgenden Auflistung fett hervorgehoben.

- QS 1, adäquate Vorgehensweise: Oberkategorie 1
- QS 2, korrektes Resultat mit richtiger Einheit: Oberkategorie 2
- QS 3, Messstrategien:
  Messwiederholung: **Oberkategorie 4**, Oberkategorie 6, **Oberkategorie 7,** Oberkategorie 12
  Mengenvergrösserung: Oberkategorie 13, **Oberkategorie 15**, **Oberkategorie 16**
- QS 4, Wahl Messinstrument: **Oberkategorie 17**

Zudem hat es im Kategoriensystem Oberkategorien zu den allgemeineren Fragen des Interviews (vgl. Oberkategorien 3, 5, 8, 9, 10, 11, 18). Diese sind im Kategoriensystem mit grauer Schrift gekennzeichnet. Die Kodierung dieser Oberkategorien wurde im Zuge der Auswertungen von FF2 nicht genutzt, kann aber für zukünftige Forschung dienen.

Anmerkungen:

Zu Beginn einer Oberkategorie wurden jeweils die Fragen aus dem Interviewleitfaden aufgeführt. Diese Fragen dienten als Orientierungshilfe bei der Kodierung.

SuS steht für *Schülerinnen und Schüler*

Kategoriensystem: Fadenaufgabe

Hinweise zur Kodierung:
→ Pro Oberkategorie jeweils nur einen Code auswählen. Es wird der Code gewählt, um welchen es bei der Antwort des Schülers oder der Schülerin *hauptsächlich* geht.
→ Wenn zu einer Oberkategorie keine Aussage gemacht wird bzw. kein Code passt: Code 777 (kann nicht beurteilt werden)

Qualitätsstandard 1 (adäquate Vorgehensweise)

Oberkategorie 1		
	Codes	**Vorgehensweise** Interviewfrage: Erkläre, wie du vorgegangen bist und warum du das so gemacht hast?
	11	Keine Aussage zur Vorgehensweise.
	12	Keine Vorgehensweise, weil Problemstellung nicht verstanden.
	13	Keine Vorgehensweise, weil keine Lust.
	14	Beschreibt Vorgehensweise aber ohne zu erklären, warum das so gemacht wurde.
	15	Ist einfach so vorgegangen, ohne speziellen Grund / es fiel nichts anderes ein / hat einfach probiert, was mit dem Material möglich ist.
	16	Ist so vorgegangen, weil gesehen wurde, dass andere SuS das so machen.
	17	Ist so vorgegangen, weil es so in der Aufgabe steht.
	18	Beschreibt eine adäquate Vorgehensweise und sagt auch, warum das so gemacht wurde (z. B. mehrmals gemessen, weil der Wert schwierig abzulesen war / den Faden mit einem Knoten an der Federwaage befestigt, weil man die Belastung herausfinden möchte, bei der ein Faden reisst / etc.)
	19	Hat es zwar so gemacht, würde es im Nachhinein aber anders machen (z. B. bei einem nächsten Mal mit doppeltem Faden messen).

Qualitätsstandard 2 (Resultat und korrekte Einheit)

Oberkategorie 2		
	Codes	**Mind. ein Wert im Toleranzbereich (TB) und mögliche Gründe, warum evtl. kein Wert in TB.** (TB: 1-Faden-Ansatz: 600-1400 g; 2-Faden-Ansatz: 1400-2800 g) Interviewfrage: Wie bist du auf dieses Resultat gekommen? ODER: Hast du ein Resultat erhalten? *Numerischer Messwert / Ergebnis kann auch erst später im Interview genannt werden.*
Kein Wert in TB. Grund: …	11	Wird nicht ersichtlich.
	12	Keine Vorgehensweise entwickelt / durchgeführt.
	13	Explizit nicht adäquate Vorgehensweise (z. B. Faden auf beiden Seiten der Federwaage befestigt (oben und unten); etc.).
	14	Schwierigkeiten bei der Durchführung (z. B. Schwierigkeiten mit dem Messinstrument (funktionierte nicht richtig), Schwierigkeiten beim Ablesen, etc.)
	15	Probleme mit Dezimalstelle (z. B. Wert ist 0.06 kg anstelle von 0.6 kg).
	16	Explizit falsche Einheit (z. B. 2 g anstelle von 2 kg)
mind. ein Wert in TB	21	Es ist mind. ein numerischer Wert für die gesuchte Grösse im TB vorhanden. Die Einheit fehlt bzw. wird im Interview auch nicht genannt.
	22	Es ist mind. ein numerischer Wert für die gesuchte Grösse im TB vorhanden. Der Wert hat die richtige Einheit.

Oberkategorie 3		
	Codes	**Glaube an die Existenz eines ‘wahren’ Werts.** Interviewfrage Max, Pascal und Cornelia haben das gleiche Experiment gemacht und folgende Belastungen erhalten: Max: **1500 g** Pascal: **1300 g** Cornelia: **1400 g** Wer hat richtig gemessen? Warum?
	11	Ich weiss es nicht.
	12	Frage kann so nicht beantwortet werden. Es können alle richtig sein (z. B. weil man nicht weiss, wie sie gemessen haben / Fäden können unterschiedlich sein / etc.)
	13	Cornelia hat richtig gemessen, ihr Ergebnis liegt in der Mitte.
	14	XXX hat richtig gemessen, weil es am nächsten bei meinem Ergebnis ist.

Qualitätsstandard 3a (Messstrategie: Messwiederholung (MW))

Oberkategorie 4		
	Codes	**Zusammenhang MW und Messgenauigkeit** Interviewfrage: Ist es bei dieser Aufgabe besser einmal oder mehrmals zu messen? Warum?
Einmal messen genügt, weil …	11	Begründung warum fehlt.
	12	… die Aufgabenstellung verlangt nicht, dass man mehrmals misst.
	13	… es wird sowieso in etwa das gleiche Ergebnis herauskommen.
	14	… man die Grösse sowieso nicht genau messen kann. Ein ungefährer Wert genügt.
	15	… wenn das Experiment genau durchgeführt wurde, dann genügt einmal messen.
Mehrmals messen ist besser, weil …	21	Begründung warum fehlt.
	22	… man hat bei dieser Aufgabe genügend Zeit.
	23	… andere, oft unlogische Begründung (z. B. je mehr man misst, desto schneller reisst der Faden).
	24	… die zu messende Grösse schwierig zu messen ist.
	25	… man sich so sicher sein kann / ein Ergebnis bestätigen kann.
	26	… man so sehen kann in welchem Bereich der Wert liegt.
	27	… man so Ausreisser erkennen und gegebenenfalls ausschliessen kann.
	28	… sich so Ungenauigkeiten ausgleichen (z. B. Faden ist nicht an allen Stellen gleich).
	29	… es so genauer wird. Ohne weitere Begründung oder nicht ausreichend verständliche Begründung, warum das so ist.
	30	… man so ein genaueres Ergebnis kriegt und einen Mittelwert berechnen kann.

Oberkategorie 5		
	Codes	**MW Allgemein** Interviewfrage: Warum wird bei *manchen* Experimenten mehr als einmal gemessen?
Es wird mehrmals gemessen, um …	11	… ein Ergebnis zu bestätigen / sich sicher zu sein / nachzuprüfen, ob man nichts falsch gemacht hat / etc.
	12	… zu sehen in welchem Bereich der Wert ungefähr ist.
	13	… Ungenauigkeiten berücksichtigen zu können.
	14	… genauer zu sein. Ohne weitere Begründung, warum das so ist.
	15	… einen Mittelwert berechnen zu können.

Oberkategorie 6		
	Codes	**Wurden bei dieser Aufgabe MW durchgeführt? Anzahl Messungen.** Interviewfrage: Wie viele Male hast du gemessen, bis du dein Endresultat hattest?
	11	1mal gemessen.
	12	2mal gemessen: Einmal mit Federwaage A und einmal mit Federwaage B. *
	13	2mal gemessen. Jeweils mit der gleichen Federwaage. *
	14	3mal gemessen. Dabei Federwaage A und B verwendet. *
	15	3mal gemessen. Dabei immer die gleiche Federwaage verwendet. *
	16	3- bis 6mal gemessen. Dabei Federwaage A und B verwendet. *
	17	3- bis 6mal gemessen. Dabei immer die gleiche Federwaage verwendet. *
	18	Mehr als 6mal gemessen. Dabei Federwaage A und B verwendet. *
	19	Mehr als 6mal gemessen. Dabei immer die gleiche Federwaage verwendet. *
* Ob immer mit der gleichen Federwaage gemessen wurde, wird evtl. erst später im Interview ersichtlich. Dann hier nachführen.		

Oberkategorie 7		
	Codes	**Warum diese Anzahl Messungen?**
		Interviewfrage: Warum hast du genau _____mal gemessen?
Keine MW (nur 1mal gemessen oder je 1mal mit Federwaage A und B), weil …	11	Begründung fehlt.
	12	… keine Zeit.
	13	… mehrmals messen nicht verlangt wurde.
	14	… sowieso in etwa das Gleiche herausgekommen wäre.
	15	… man kann es sowieso nicht genau messen. Ein ungefährer Wert genügt.
	16	… genau gearbeitet wurde. MW waren nicht nötig.
	17	… die Instrumente verglichen wurden (eine Messung mit Federwaage A und eine mit Federwaage B).
MW (mind. 2mal mit gleichem Messinstrument gemessen), weil …	21	Begründung fehlt.
	22	… einfach so / es gibt keinen Grund dafür.
	23	… man nicht mehr genügend Zeit für noch mehr Messungen hatte.
	24	… man ausreichend Zeit für so viele Messungen hatte.
	25	… man so viele Messungen als eine gute Anzahl / als ausreichend erachtete.
	26	… zu viele Messungen sind auch nicht gut. Verwirren nur.
	27	… in den Naturwissenschaften immer so oft gemessen wird (z. B. immer 3mal).
	28	… es bei den ersten Messungen noch Schwierigkeiten gab / ich nachher wusste, wie es geht.
	29	…. man sich so sicher sein kann / ein Ergebnis bestätigt wurde.
	30	… man dann sehen konnte, in welchem Bereich die Werte waren.
	31	… es so genauer wird. Ohne weitere Begründung warum.
	32	… man so mit beiden Federwaagen ausreichend viele Messungen hatte.
	33	… man so Ungenauigkeiten ausgleichen kann.
	34	… es so genauer wird und man einen Mittelwert berechnen kann.

<table>
<tr><td colspan="3">Oberkategorie 8</td></tr>
<tr><td></td><td>Codes</td><td>Allgemeine Fragen zum MW</td></tr>
<tr><td></td><td colspan="2">Julian hat dreimal hintereinander gemessen. Die Belastungen betrugen dabei:
<table><tr><td>Messung Nr.</td><td>Belastungen</td></tr><tr><td>1</td><td>1500 g</td></tr><tr><td>2</td><td>1400 g</td></tr><tr><td>3</td><td>1500 g</td></tr></table>Wie kann er mit diesen Resultaten zu einem Endresultat kommen?</td></tr>
<tr><td></td><td>11</td><td>Keine Ahnung wie man zu einem Schlussergebnis kommen kann.</td></tr>
<tr><td></td><td>12</td><td>Wählt 1500 g, weil dies der letzte Wert ist.</td></tr>
<tr><td></td><td>13</td><td>Nimmt wiederholten Wert (also 1500 g).</td></tr>
<tr><td></td><td>14</td><td>Wählt einen Wert, weil dieser am nächsten beim persönlich gemessenen Wert liegt.</td></tr>
<tr><td></td><td>15</td><td>Wählt einen mittleren Wert (Wert zwischen 1400 und 1500 g, also 1450 g).</td></tr>
<tr><td></td><td>16</td><td>Würde einen Mittelwert berechnen. Im Interview wird aber nicht klar, ob er / sie weiss wie das geht.</td></tr>
<tr><td></td><td>17</td><td>Versucht von allen Werten einen Mittelwert zu berechnen. Vorgehensweise ist aber nicht korrekt.</td></tr>
<tr><td></td><td>18</td><td>Berechnet von allen Werten korrekt einen Mittelwert.</td></tr>
<tr><td colspan="3">Oberkategorie 9</td></tr>
<tr><td></td><td colspan="2">Interviewfrage: Wenn man noch mehr MW macht, zum Beispiel zehn Messungen anstellen von drei, wird dann das Ergebnis genauer oder bleibt es gleich genau. Warum?</td></tr>
<tr><td></td><td>11</td><td>Unlogische Begründung (z. B. Ergebnis wird immer grösser).</td></tr>
<tr><td></td><td>12</td><td>Ergebnis kann auch ungenauer werden. Keine weitere Begründung / unvollständige oder unklare Begründung.</td></tr>
<tr><td></td><td>13</td><td>Ergebnis kann auch ungenauer werden: Man kriegt immer mehr und unterschiedliche Zahlen.</td></tr>
<tr><td></td><td>14</td><td>Ergebnis bleibt gleich genau. Keine weitere Begründung / unvollständige oder unklare Begründung.</td></tr>
<tr><td></td><td>15</td><td>Ergebnis bleibt gleich genau. Werte werden sich einfach wiederholen / Resultat wird einfach bestätigt.</td></tr>
<tr><td></td><td>16</td><td>Ergebnis wird genauer. Keine weitere Begründung / unvollständige oder unklare Begründung.</td></tr>
<tr><td></td><td>17</td><td>Ergebnis wird genauer. Man sieht besser, in welchem Bereich die Werte liegen.</td></tr>
<tr><td></td><td>18</td><td>Ergebnis wird genauer. Ungenauigkeiten gleichen sich aus.</td></tr>
<tr><td></td><td>19</td><td>Ergebnis wird genauer. Man kann sich sicherer sein.</td></tr>
<tr><td></td><td>20</td><td>Ergebnis wird genauer, weil man so mehr Stellen nach dem Dezimalpunkt kriegt.</td></tr>
</table>

<table>
<tr><td></td><td>21</td><td>Ergebnis wird genauer. Weil man einen Mittelwert berechnen kann / ein Mittelwert von mehr Zahlen ist genauer / etc.</td></tr>
<tr><td colspan="3">Oberkategorie 10</td></tr>
<tr><td></td><td colspan="2">Iva hat beim Experiment folgende Ergebnisse erhalten:
<table><tr><td>Messung Nr.</td><td>Belastungen</td></tr><tr><td>1</td><td>1300 g</td></tr><tr><td>2</td><td>1400 g</td></tr><tr><td>3</td><td>1300 g</td></tr><tr><td>4</td><td>2500 g</td></tr><tr><td>5</td><td>1500 g</td></tr></table>Wie kommt sie zu einem Endresultat?</td></tr>
<tr><td rowspan="7">Erkennt Ausreisser nicht und …</td><td>11</td><td>… nimmt 1500 g, weil dies der letzte Messwert ist.</td></tr>
<tr><td>12</td><td>… wählt einen Wert, der am nächsten beim persönlich gemessenen Wert liegt.</td></tr>
<tr><td>13</td><td>… nimmt 1300 g, weil dieser Wert bestätig wurde.</td></tr>
<tr><td>14</td><td>… wählt einen mittleren Wert (Wert zwischen 1300 und 2500 g).</td></tr>
<tr><td>15</td><td>… würde von allen Messungen den Mittelwert berechnen. Im Interview wird nicht klar, ob er / sie weiss, wie das geht.</td></tr>
<tr><td>16</td><td>… versucht von allen Messungen den Mittelwert zu berechnen. Vorgehensweise zur Berechnung ist aber nicht korrekt.</td></tr>
<tr><td>17</td><td>… berechnet von allen Messungen den Mittelwert. Vorgehensweise zur Berechnung ist korrekt.</td></tr>
<tr><td rowspan="7">Erkennt Ausreisser, behaltet diesen bei und …</td><td>21</td><td>… nimmt 1500 g, weil dies der letzte Messwert ist.</td></tr>
<tr><td>22</td><td>… wählt einen Wert, der am ehesten mit dem persönlichen Wert übereinstimmt.</td></tr>
<tr><td>23</td><td>… nimmt 1300 g, weil dieser Wert bestätig wurde.</td></tr>
<tr><td>24</td><td>… wählt einen mittleren Wert (Wert zwischen 1300 und 2500 g).</td></tr>
<tr><td>25</td><td>… würde von allen Messungen den Mittelwert berechnen. Im Interview wird nicht klar, ob er / sie weiss, wie das geht.</td></tr>
<tr><td>26</td><td>… versucht von allen Werten einen Mittelwert zu berechnen. Vorgehensweise zur Berechnung ist aber nicht korrekt.</td></tr>
<tr><td>27</td><td>… berechnet von allen Werten einen Mittelwert. Vorgehensweise zur Berechnung ist korrekt.</td></tr>
<tr><td rowspan="6">Stark abweichende Messung wiederholen und …</td><td>31</td><td>… dann schauen in welchem Bereich die Werte liegen.</td></tr>
<tr><td>32</td><td>… dann schauen ob sich ein Wert wiederholt.</td></tr>
<tr><td>33</td><td>… dann einen mittleren Wert wählen.</td></tr>
<tr><td>34</td><td>… dann von allen Werten einen Mittelwert berechnen. Im Interview wird nicht klar, ob er / sie weiss, wie das geht.</td></tr>
<tr><td>35</td><td>… dann versuchen von allen Werten einen Mittelwert zu berechnen. Vorgehensweise zur Berechnung kann aber nicht korrekt genannt werden.</td></tr>
<tr><td>36</td><td>… dann von allen Werten einen Mittelwert berechnen. Vorgehensweise zur Berechnung kann korrekt genannt werden.</td></tr>
</table>

Schliesst Ausreisser aus und …	41	… nimmt dann einen mittleren Wert (Wert zwischen 1300 und 1500 g, also 1400 g).
	42	… gibt dann ein Spektrum an (1300 -1500 g).
	43	… würde von den restlichen Werten einen Mittelwert berechnen. Im Interview wird nicht klar, ob er / sie weiss, wie das geht.
	44	… versucht von den restlichen Werten einen Mittelwert zu berechnen. Vorgehensweise zur Berechnung ist aber *nicht korrekt*.
	45	… berechnet von den restlichen Werten einen Mittelwert. Vorgehensweise zur Berechnung ist *korrekt*.

Oberkategorie 11

Fabian und Alexandra haben beide viermal gemessen. Sie haben folgende Belastungen erhalten:

Fabian	
Messung Nr.	Belastung
1	**1400 g**
2	**1500 g**
3	**1300 g**
4	**1400 g**

Alexandra	
Messung Nr.	Belastung
1	**1800 g**
2	**1400 g**
3	**1100 g**
4	**1300 g**

Welchen Resultaten kannst du mehr vertrauen? Warum?

11	xxx ist vertrauenswürdiger. Unlogische Begründung (z. B. habe auch festgestellt, dass der Faden immer schneller reisst).
12	Fabian / Alexandra ist vertrauenswürdiger. Ich habe ähnliche Werte gekriegt.
13	Beide sind gleich vertrauenswürdig. Man weiss nicht, wie sie gemessen haben.
14	Beide sind gleich vertrauenswürdig. Beide haben den gleichen Mittelwert.
15	Fabian ist vertrauenswürdiger, weil er 2mal das gleiche Resultat hat.
16	Fabian ist vertrauenswürdiger, da die Werte näher beieinander sind ODER: Fabian ist vertrauenswürdiger, da Alexandras Werte zu unterschiedlich sind.

Oberkategorie 12

Codes	**MW, Bedingungen konstant halten**
	Interviewfrage: Wenn mehrmals gemessen wird, muss man dann beides Mal das gleiche Messinstrument nehmen oder kann man bei jeder Messung ein anderes nehmen. Warum?
11	Unverständliche Aussage.
12	Kann beide Federwaagen nehmen. Keine weiteren Angaben oder unverständliche Angaben.
13	Kann beide Federwaagen nehmen. Beide messen die Belastung.
14	Kann eigentlich beide Federwaagen nehmen / hätte z. B. mit beiden Federwaagen einmal ausprobieren können, zum Vergleichen / etc.
15	Man sollte immer die gleiche Federwaage nehmen. Keine weiteren oder unverständliche Angaben.

	16	Man sollte immer die gleiche Federwaage nehmen, weil sie messen unterschiedlich (z. B. Aussagen wie: «die eine Federwaage reagiert schneller als die andere», «die eine zeigt auch 20 g / 50 g Schritte an und die andere nicht», etc.).

Qualitätsstandard 3b (Messstrategie: Messen mit einer (grossen) Menge (GM))

Oberkategorie 13		
	Codes	**Einstieg GM**
	Interviewfrage: Spielt es bei dieser Aufgabe eine Rolle, ob man den Faden wie in Bild 1 (links) oder wie in Bild 2 (rechts) befestigt? Warum?  Werden sich die Ergebnisse unterscheiden, je nachdem ob man wie in Bild 1 oder wie in Bild 2 misst?	
	11	Unverständliche Aussage.
	12	Es spielt *keine* Rolle ob der Faden wie in Bild 1 oder wie in Bild 2 befestigt wird. Keine weitere oder unverständliche Aussage.
	13	Es spielt *keine* Rolle. Die Belastung auf den Faden ist immer gleich.
	14	Es spielt *keine* Rolle. Solange man bei Bild 2 zurückrechnet.
	15	Es spielt *eine* Rolle ob der Faden wie in Bild 1 oder wie in Bild 2 befestigt wird. Die Ergebnisse werden sich unterscheiden. Keine weitere Aussage oder unverständliche Aussage.
	16	Es spielt *eine* Rolle. Die Ergebnisse werden sich unterscheiden. Bei Bild 1 braucht es eine grössere Belastung bis der Faden reisst, weil beim Knopf mehrere Fäden sind.
	17	Es spielt *eine* Rolle. Die Ergebnisse werden sich unterscheiden. Bei Bild 2 braucht es eine grössere Belastung bis der Faden reisst, da der Faden doppelt ist.

Oberkategorie 14	
	Code 333 (Oberkategorie 14 ist bei der Fadenaufgabe nicht besetzt).

Oberkategorie 15		
	Codes	**Zusammenhang zwischen GM und Messgenauigkeit** Interviewfrage: Ist es bei dieser Aufgabe besser so (Bild 1, Oberkategorie 13) oder so (Bild 2, Oberkategorie 13) zu messen? Warum?
Keine Aussage was besser ist	11	Keine weitere Begründung oder unverständliche Aussage.
	12	… man müsste es ausprobieren / ich weiss es nicht.
	13	… man kann wie in Bild 1 oder wie in Bild 2 messen. Solange man bei Bild 2 auf einen Faden zurückrechnet.
Es ist besser wie in Bild 1 (mit einem Faden) zu messen, weil…	21	Keine weitere Begründung / unvollständige Begründung / unklare Begründung.
	22	… es steht so in der Aufgabe.
	23	… es steht so in der Aufgabe. Theoretisch würde es aber auch wie in Bild 2 gehen (evtl. Ergänzung: Dann müsste die Aufgabe aber anders lauten).
	24	… so die Gefahr nicht besteht, dass man sich verrechnet.
Es ist besser wie in Bild 2 (mit mehr als 1 Faden) zu messen, weil…	31	Keine weitere Begründung / unvollständige Begründung / unklare Begründung.
	32	… die Belastung, bei welcher 1 Faden reisst, ist schwierig zu messen.
	33	… man braucht so keinen Knoten zu machen.
	34	… man kann so schauen, welcher Faden zuerst reisst und ob sie gleichzeitig reissen.
	35	… die Belastung, bei welcher 1 Faden reisst, zu messen wird sehr ungenau, mit mehr als 1 Faden wird es genauer.

Oberkategorie 16		
	Codes	**Begründung warum bei dieser Aufgabe mit einer GM gemessen wurde.** Interviewfrage: Mit wie vielen Fäden hast du aufs Mal gemessen? Warum hast du genau _____ Fäden aufs Mal genommen?
	11	… erkennt nicht, dass mit einer Menge gemessen wurde (z. B. sagt, hat mit 1 Faden gemessen, obwohl es eigentlich 2 Fäden waren (z. B. mit Schlaufe gemessen))
Mit 1 Faden gemessen, weil …	21	Keine weitere Begründung.
	22	… es so in Aufgabe steht.
	23	… es so einfacher ist: Man muss nicht zurückrechnen.
Mit 2 Fäden gemessen, weil …	31	Keine weitere Begründung.
	32	… man muss so keinen Knoten machen.
	33	… wenn man nur mit 1 Faden misst, dann reisst es beim Knoten.
	34	… die Belastung für 1 Faden zu messen schwierig ist (geht sehr schnell).
	35	… es so genauer wird.
Mehrmaliges Umwickeln des Fadens, weil…	41	Keine weitere Begründung.
	42	… man muss so keinen Knoten machen.
	43	… die Belastung für 1 Faden zu messen schwierig ist (geht sehr schnell).
	44	… es so genauer wird.
Mit 1 Faden und einer Menge gemessen, weil …	51	Keine weitere Begründung.
	52	… man merkte, dass die Belastung mit 1 Faden schwierig zu messen ist.
	53	… man so vergleichen kann.
	54	… aus Neugierde.

Qualitätsstandards 4 (Messinstrument)

Oberkategorie 17		
	Codes	**Zusammenhang Wahl Messinstrument und Messgenauigkeit** Interviewfrage: Hast du Federwaage A oder Federwaage B genommen? Warum hast du dich für diese Federwaage entschieden?
Hat Federwaage A genommen, weil …	11	Keine Begründung.
	12	… diese übersichtlicher ist.
	13	… diese besser zu bedienen ist.
	14	… ein Messbereich von 2.5 kg ausreicht.
	15	… diese in Gramm anzeigt und Federwaage B in Kilogramm.
	16	… diese genauer ist. Ohne weitere Begründung.
	17	… diese genauer ist und vom Messbereich ausreicht.
	18	… diese genauer ist. Mit Verweis auf die feinere Skala (z. B. Federwaage A kann auf 20 g bzw. 50 g genau messen / Federwaage B misst nur auf 100 g genau / etc.) *
Nimmt beide Federwaagen, weil …	21	Keine Begründung.
	22	… zum Vergleichen (z. B. auch: Am Anfang mit beiden ausprobiert, dann Federwaage A oder B).
Nimmt Federwaage B, weil …	31	Keine Begründung.
	32	… diese übersichtlicher ist.
	33	… diese besser zu bedienen ist.
	34	… Federwaage B mehr Belastung aushält, ich hatte Angst, dass Federwaage A kaputt geht.
* Auch wenn Skala falsch interpretiert wurde (z. B. Milligramm).		

Genauigkeit der Angabe des Ergebnisses:

Oberkategorie 18		
	Codes	**Stellen nach dem Dezimalpunkt** Interviewfrage: Ava hat als Endresultat eine Belastung von 1433.33 g aufgeschrieben. Ist es sinnvoll so viele Stellen nach dem Punkt anzugeben oder würden auch weniger Stellen nach dem Punkt ausreichen? Warum?
So viele Stellen sind gut, weil …	11	Keine weitere / unverständliche / nicht klare Begründung.
	12	… es so genauer ist / man sieht, dass es ein genaues Ergebnis ist.
	13	… man hat die Stellen sowieso, dann kann man sie auch angeben.
Weniger Stellen reichen aus, weil …	21	Keine weitere / unverständliche / nicht klare Begründung.
	22	… man kann es sowieso nicht so genau sagen / 0.33 g bzw. 0.03 g ist so wenig, das kann man vernachlässigen.
Kann man so nicht sagen, weil …	31	Keine weitere / unverständliche / nicht klare Begründung.
	32	… es darauf ankommt, für was das Ergebnis gebraucht wird (z. B. wenn mit dem Ergebnis weitergerechnet wird, sind so viele Stellen gut; für die Schule würden weniger Stellen ausreichen; für die Wissenschaft braucht es so viele Stellen; etc.)

Abschliessende Fragen des Interviews:

- **Interviewfrage: Spielt es für die Genauigkeit eine Rolle ob Federwaage A oder Federwaage B verwendet wird?**
  → **Falls im Interview zu dieser Frage Antworten vorhanden sind: Abgleichen mit Oberkategorie 17.**
  - Wenn vorhin kein Verweis auf genauere Skala und nun schon: Bei Oberkategorie 17 Code ändern.
  - Falls hier ein anderer Grund für die Wahl des Messinstruments genannt wird: Bei Oberkategorie 17 besseren Code (Codes mit Federwaage A) oder höheren Code in einer Subkategorie wählen.

- **Interviewfrage: Wenn man mehrmals misst, wird dann bei dieser Aufgabe das Resultat genauer? Warum?**
  → **Falls im Interview zu dieser Frage Antworten vorhanden sind: Abgleichen mit Oberkategorie 4.**
  - Wenn vorhin genannt wurde einmal messen genügt und nun erkannt wird, dass mehrmals messen besser ist: Bei Oberkategorie 4 Code ändern, höherer Code zählt.
  - Falls vorhin nicht begründet wurde warum mehrmals messen genauer ist und nun schon: Bei Oberkategorie 4 Code ändern, höherer Code zählt.
  - Falls hier ein anderer Grund für Messwiederholung genannt wird: Bei Oberkategorie 4 Code ändern, höherer Code zählt.

- **Interviewfrage: Wenn man mit mehreren Fäden aufs Mal misst, wird dann das Resultat genauer? Warum?**
  → **Falls im Interview zu dieser Frage Antworten vorhanden sind: Abgleichen mit Oberkategorie 15**
  - Wenn vorhin genannt wurde, dass mit 1 Faden besser ist und nun erkannt wird, dass besser mit einer Menge gemessen wird: Bei Oberkategorie 15 Code ändern, höherer Code zählt.
  - Falls vorhin nicht begründet wurde warum besser mit mehr als 1 Faden gemessen wird und nun schon: Bei Oberkategorie 15 Code ändern, höherer Code zählt.
  - Falls hier ein anderer Grund für das Messen mit mehr als einem Faden genannt wird: Bei Oberkategorie 15 Code ändern, höherer Code zählt.

C Expertenrating zur Einschätzung der kategorisierten Schüleraussagen

Die von den Schülerinnen und Schülern im Interview gezeigten kategorisierten Schüleraussagen im Bereich der intendierten Konzepte (QS 3 und QS 4; vgl. auch Kategoriensystem, Anhang - Teil B) wurden ins Expertenrating aufgenommen und den Expertinnen und Experten zur Einschätzung vorgelegt. Die Expertinnen und Experten schätzten die Aussagen im Hinblick darauf ein, inwiefern sie den intendierten Konzepten entsprechen und somit Hinweise darüber vorliegen, dass die Aufgaben kognitiv valide Schlüsse bezüglich der experimentellen Kompetenzen von Schülerinnen und Schülern im Bereich des naturwissenschaftlichen Messens zulassen.
In der Folge wird das Expertenrating aufgeführt. Hierbei werden die Antwortfenster des PDF-Formulars nicht ersichtlich. In Unterkapitel 8.2.2 wurde der Aufbau des Ratings beschrieben, wobei in den Abbildungen von den Ausschnitten aus dem Rating auch die Antwortfenster des PFD-Formulars ersichtlich werden.

Anmerkungen:

S.	steht für *Schülerin* oder *Schüler*
S.-Aussagen	steht für *Schüleraussagen*

Die im Expertenrating aufgeführten Aufgaben wurden in Unterkapitel 6.4.1 beschrieben und im Rahmen des ExKoNawi-Projekts entwickelt (vgl. z. B. Gut et al., 2017; Metzger et al., 2014; Bonetti, in Vorbereitung).

Die Fotos von den Experimentiermaterialien sind von Bonetti.

Expertenrating

Die folgende Umfrage findet im Rahmen des Projekts ExKoNawi (**Ex**perimentelle **Ko**mpetenzen in den **Na**tur**wi**ssenschaften) statt. Bei den Erhebungen haben Schülerinnen und Schüler der Sekundarstufe I Aufgaben mit Realexperimenten durchgeführt und wurden anschliessend interviewt. Im Rahmen dieser Auswertung liegt der Fokus auf den Aufgaben des Problemtyps «Naturwissenschaftliches Messen mit vorgegebenen Instrumenten». Das Ziel dieser Aufgaben ist es, eine Grösse mit vorgegebenen Instrumenten so genau wie möglich zu messen. Dafür stehen den Schülerinnen und Schülern jeweils zwei verschiedene Messinstrumente zur Verfügung, die unterschiedlich genau messen. Ziel ist, dass die Schülerinnen und Schüler das genauere Messinstrument wählen, Messwiederholungen durchführen und evtl. mit einer Menge messen, um ein möglichst genaues Ergebnis zu erzielen.
Auf den folgenden Seiten werden kategorisierte Aussagen der Schülerinnen und Schüler aus den Interviews aufgeführt. Gerne möchte ich Sie bitten, die Aussagen auf einer Skala einzuschätzen. Zum Bearbeiten benötigen Sie circa 60 Minuten.

Ich danke Ihnen herzlich für Ihr Engagement und Ihre Teilnahme!

Freundliche Grüsse
Livia Murer

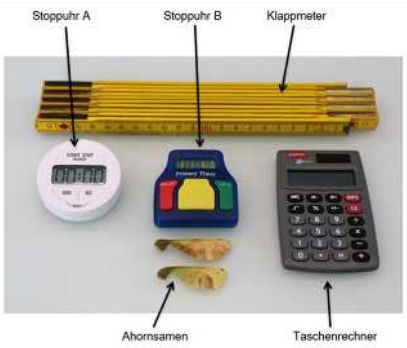
Ahornaufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, wie lange ein Ahornsamen drehend braucht, um die Strecke einer Pulthöhe zu fallen.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Stoppuhr A: Misst auf Sekunden genau.
- Stoppuhr B: Misst auf Hundertstelsekunden genau.
- Ein Ahornsamen braucht drehend circa 0.3 s bis 1.6 s, um die Strecke einer Pulthöhe hinunterzufallen.



	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (z. B. 1.5- oder 2-fache Pulthöhe) zu messen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Die Zeit für eine Menge kann besser gestoppt werden und die Reaktionszeit zum «Start»- und «Stopp»-Drücken hat weniger Einfluss auf das Messergebnis.				

Möglichkeit für Rückmeldungen zur Einschätzung:

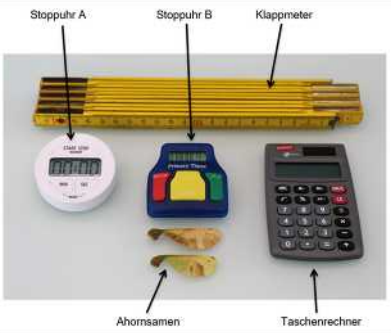
Ahornaufgabe – 2

Die Schülerinnen und Schüler sollen herausfinden, wie lange ein Ahornsamen drehend braucht, um die Strecke einer Pulthöhe zu fallen.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Stoppuhr A: Misst auf Sekunden genau.
- Stoppuhr B: Misst auf Hundertstelsekunden genau.
- Ein Ahornsamen braucht drehend circa 0.3 s bis 1.6 s um die Strecke einer Pulthöhe hinunterzufallen.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Ahornaufgabe geht es um das Konzept: Wenn mit mehr als einer Pulthöhe gemessen wird (z. B. mit 1.5- oder 2-facher Pulthöhe), dann erhöht dies die Messgenauigkeit. Wenn mit mehr als einer Pulthöhe gemessen wird, hat die Reaktionszeit zum «Start»- und «Stopp»-Drücken weniger Einfluss auf das Endergebnis. Zudem kann das Experiment genauer durchgeführt werden: Mit 1.5- oder 2-facher Pulthöhe dreht der Ahornsamen besser.

Zu prüfen wäre also, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der gemessenen Strecke und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe mit der doppelten Pulthöhe gemessen, weil es genauer wird» oder «Ich habe mit der doppelten Pulthöhe gemessen, weil dann der Ahornsamen besser dreht und es so genauer wird») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nicht über die gemessene Strecke nachgedacht, weil sie keinen Einfluss auf die Messgenauigkeit hat»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe mit der doppelten Pulthöhe gemessen, weil ich mich bewegen und aufstehen wollte», weil hier die gewählte Strecke vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. begründet nicht, ob es bei dieser Aufgabe besser ist, mit einer oder mit mehr als einer Pulthöhe zu messen.							
S.: «Es ist bei dieser Aufgabe besser, mit einer Pulthöhe zu messen.» S. begründet nicht warum.							
S. hat mit einer beliebigen Höhe gemessen (z. B. 1 m oder 1.2 m). S.: «Weil ich die Pulthöhe so geschätzt habe.»							
S.: «Es ist bei dieser Aufgabe besser, mit einer Pulthöhe zu messen, weil es so in der Aufgabe steht.»							
S.: «Es ist bei dieser Aufgabe besser, mit einer Pulthöhe zu messen, weil so nicht die Gefahr besteht, dass man sich verrechnet.»							
S.: «Es ist bei dieser Aufgabe besser, mit einer Pulthöhe zu messen. So muss man die Strecke nicht mit dem Klappmeter abmessen.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Es ist bei dieser Aufgabe besser, mit mehr als einer Pulthöhe zu messen. Die Zeit für eine Pulthöhe zu messen ist schwierig, das wird sehr ungenau.»							
S.: «Es ist bei dieser Aufgabe besser, mit mehr als einer Pulthöhe zu messen, weil der Ahornsamen bei einer Pulthöhe kaum drehte.»							
S.: «Es ist bei dieser Aufgabe besser, mit mehr als einer Pulthöhe zu messen, weil die Zeit für eine Pulthöhe sehr ungenau wird. Mit mehr als einer Pulthöhe wird es genauer.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

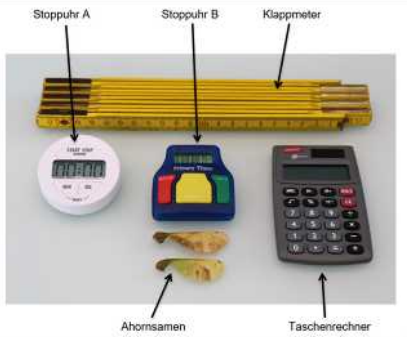
Ahornaufgabe – 3

Die Schülerinnen und Schüler sollen herausfinden, wie lange ein Ahornsamen drehend braucht, um die Strecke einer Pulthöhe zu fallen.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Stoppuhr A: Misst auf Sekunden genau.
- Stoppuhr B: Misst auf Hundertstelsekunden genau.
- Ein Ahornsamen braucht drehend circa 0.3 s bis 1.6 s um die Strecke einer Pulthöhe hinunterzufallen.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Ahornaufgabe geht es um das Konzept: Wenn das genauere Messinstrument (hier Stoppuhr B) verwendet wird, dann erhöht dies die Messgenauigkeit.

Zu prüfen wäre also bei der Ahornaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen dem gewählten Messinstrument und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Stoppuhr B genommen, weil diese genauer ist») als auch falsche Bezüge zum Konzept (z. B. «Man kann beide Stoppuhren nehmen. Die Wahl der Stoppuhr hat keinen Einfluss auf die Messgenauigkeit.»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe Stoppuhr B genommen, weil diese farbig ist», weil hier die Wahl des Messinstruments vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe Stoppuhr B genommen, weil Stoppuhr A nicht richtig funktionierte.»							
S.: «Ich habe Stoppuhr B genommen, weil diese genauer ist.» Ohne weitere Begründung.							
S.: «Ich habe Stoppuhr B genommen, weil diese genauer ist.» S. verweist auf die feinere Skala (z. B. Stoppuhr B zeigt mehr Stellen an, Stoppuhr B misst auch Hundertstelsekunden).							

Möglichkeit für Rückmeldungen zur Einschätzung:

Bohnenaufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, wie schwer eine einzelne Bohne ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Waage A: 10 g-Skalierung.
- Waage B: 2 g-Skalierung.
- Die Masse einer getrockneten Bohne beträgt circa 0.3 g bis 1 g.

	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Messwiederholung bedeutet bei dieser Aufgabe mehrmals hintereinander mit einer bestimmten Anzahl Bohnen zu messen und dabei die Bohnen auszutauschen.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (z. B. 20 Bohnen auf einmal) zu messen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Die Masse einer einzelnen Bohne kann mit den gegebenen Messinstrumenten kaum ermittelt werden und wird sehr ungenau. Wenn mit einer Menge gemessen wird (z. B. 20 Bohnen auf einmal), wird das Ergebnis genauer und verschiedene Bohnengrössen werden berücksichtigt.				

Möglichkeit für Rückmeldungen zur Einschätzung:

Bohnenaufgabe – 2

Die Schülerinnen und Schüler sollen herausfinden, wie schwer eine einzelne Bohne ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Waage A: 10 g-Skalierung.
- Waage B: 2 g-Skalierung.
- Die Masse einer getrockneten Bohne beträgt circa 0.3 g bis 1 g.

Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Bohnenaufgabe geht es um das Konzept: Wenn mit einer Menge gemessen wird (z. B. 10 Bohnen auf einmal), dann erhöht dies die Messgenauigkeit. Dabei gilt, dass je grösser die Menge (z. B. 20 oder 50 Bohnen auf einmal), desto grösser die Messgenauigkeit. Es kann mit einer Menge gemessen werden, weil das Experiment so genauer durchführbar ist (z. B. mit 10 Bohnen auf einmal, weil man es genauer auf der Skala der Waage ablesen kann) oder weil es tatsächlich genauer wird (z. B. 50 Bohnen auf einmal, weil man so unterschiedliche Bohnengrössen berücksichtigen kann). Diese Gründe für das Messen mit einer Menge widerspiegeln dabei unterschiedliche Niveaus der S.

Zu prüfen wäre also bei der Bohnenaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der Anzahl gemessener Bohnen und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe 20 Bohnen auf einmal genommen, damit die Messung genauer wird» oder «Ich habe 10 Bohnen auf einmal genommen, weil ich es so genauer bei der Waage ablesen kann») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nicht über die Anzahl Bohnen nachgedacht, weil diese keinen Einfluss auf die Genauigkeit hat»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe 5 Bohnen auf einmal genommen, weil ich nicht so viele Bohnen abzählen wollte», weil hier die Auswahl der Anzahl Bohnen vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe mit einer und mit mehreren Bohnen gemessen, zum Vergleichen.»							
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil bei einer Bohne die Waage noch nichts anzeigt, mit mehreren Bohnen kann man es besser auf der Skala der Waage ablesen.»							
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil die Masse für eine Bohne zu bestimmen sehr ungenau wird. Mit mehreren Bohnen wird es genauer.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil Bohnen unterschiedliche Massen haben. Wenn man mit mehreren Bohnen auf einmal misst wird es genauer, weil man eher eine durchschnittliche Bohne kriegt.»							
S.: «Ich habe mit mehreren Bohnen auf einmal gemessen, weil der Plastikbecher auch eine Masse hat. Je mehr Bohnen man nimmt, desto eher kann die Masse des Bechers vernachlässigt werden.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

Bohnenaufgabe – 3

Die Schülerinnen und Schüler sollen herausfinden, wie schwer eine einzelne Bohne ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Waage A: 10 g-Skalierung.
- Waage B: 2 g-Skalierung.
- Die Masse einer getrockneten Bohne beträgt circa 0.3 g bis 1 g.

Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Bohnenaufgabe geht es um das Konzept: Wenn das genauere Messinstrument (hier Waage B) verwendet wird, dann erhöht dies die Messgenauigkeit.

Zu prüfen wäre also bei der Bohnenaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen dem gewählten Messinstrument und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Waage B genommen, weil diese genauer ist») als auch falsche Bezüge zum Konzept (z. B. «Man kann beide Waagen nehmen. Die Wahl der Waage hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe Waage B genommen, weil ich die Skala von Waage A nicht verstanden habe», weil hier die Auswahl des Messinstruments vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

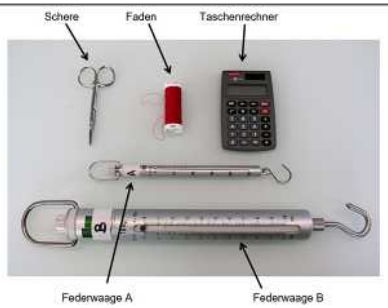
	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe Waage B genommen, weil diese übersichtlicher ist. Man kann es besser ablesen.»							
S.: «Ich habe Waage B genommen, weil diese genauer ist.» Mit Verweis auf die feinere Skala (z. B. Waage B zeigt mehr Stellen an, zeigt auf 2 g genau an).							

Möglichkeit für Rückmeldungen zur Einschätzung:

Fadenaufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, bei welcher Belastung ein Nähfaden reisst.

Dazu erhalten sie das abgebildete Material.



Hinweise für Expertinnen und Experten

- Federwaage A: Maximale Belastung 2500 g; die Skala ist auf 20 g genau beschriftet.
- Federwaage B: Maximale Belastung 10 kg; die Skala ist auf 100 g genau beschriftet.
- Ein Nähfaden reisst bei einer Belastung von circa 600 g bis 1400 g.

	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (z. B. doppeltem Faden) zu messen, um ein möglichst genaues Ergebnis zu erhalten.				

Hinweis: Anstelle davon, dass mit einem Faden gemessen und ein Knoten gemacht wird (Bild 1), kann z. B. auch mit doppeltem Faden gemessen werden (Bild 2). So kann genauer abgelesen werden, bei welcher Belastung der Faden reisst und es muss kein Knoten gemacht werden.

Bild 1

Bild 2

Möglichkeit für Rückmeldungen zur Einschätzung:

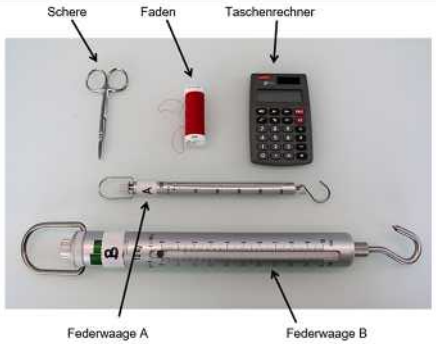
Fadenaufgabe – 2

Die Schülerinnen und Schüler sollen herausfinden, bei welcher Belastung ein Nähfaden reisst.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Federwaage A: Maximale Belastung 2500 g; die Skala ist auf 20 g genau beschriftet.
- Federwaage B: Maximale Belastung 10 kg; die Skala ist auf 100 g genau beschriftet.
- Ein Nähfaden reisst bei einer Belastung von circa 600 g bis 1400 g.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Fadenaufgabe geht es um das Konzept: Wenn mit mehr als einem Faden gemessen wird (z. B. mit doppeltem Faden), dann erhöht dies die Messgenauigkeit. Die Messgenauigkeit wird dabei erhöht, indem genauer abgelesen werden kann, bei welcher Belastung der Faden reisst.

Zu prüfen wäre also, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der Anzahl gemessener Fäden (z. B. doppelter Faden) und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe mit doppeltem Faden gemessen, weil es genauer wird») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nicht darüber nachgedacht, ob ich mit einem oder dem doppelten Faden messe, weil es keinen Einfluss auf die Messgenauigkeit hat»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe nur mit einem Faden gemessen, weil ich die Federwaage nicht kaputt machen wollte», weil hier die Anzahl gemessener Fäden vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. begründet nicht, ob es bei dieser Aufgabe besser ist, mit einem oder mehr als einem Faden auf einmal zu messen.							
S.: «Es ist bei dieser Aufgabe besser, mit einem Faden zu messen.» S. begründet nicht warum.							
S.: «Es ist bei dieser Aufgabe besser, mit einem Faden zu messen, weil es so in der Aufgabe steht.»							
S.: «Es ist bei dieser Aufgabe besser, mit 2 Fäden auf einmal zu messen (Bild), weil man so keinen Knoten machen muss.» **Bild** 							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Es ist bei dieser Aufgabe besser, mit 2 Fäden auf einmal zu messen. So kann man schauen welcher Faden zuerst reisst oder ob beide gleichzeitig reissen.»							
S.: «Ich habe mit einem und mit mehreren Fäden gemessen. Zum Vergleichen, aus Neugierde.»							
S. «Es ist besser mit 2 Fäden auf einmal zu messen. Wenn man nur mit einem Faden misst, dann reisst es eher beim Knoten.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

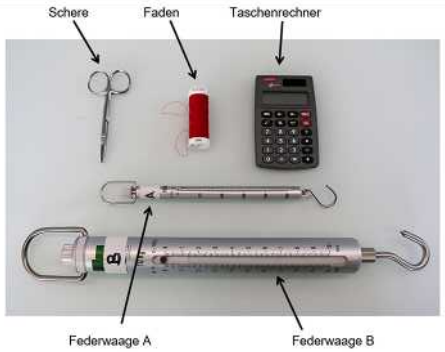
Fadenaufgabe – 3

Die Schülerinnen und Schüler sollen herausfinden, bei welcher Belastung ein Nähfaden reisst.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Federwaage A: Maximale Belastung 2500 g; die Skala ist auf 20 g genau beschriftet.
- Federwaage B: Maximale Belastung 10 kg; die Skala ist auf 100 g genau beschriftet.
- Ein Nähfaden reisst bei einer Belastung von circa 600 g bis 1400 g.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Fadenaufgabe geht es um das Konzept: Wenn das genauere Messinstrument (hier Federwaage A) verwendet wird, dann erhöht dies die Messgenauigkeit.

Zu prüfen wäre also bei der Fadenaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen dem gewählten Messinstrument und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Federwaage A genommen, weil diese genauer ist») als auch falsche Bezüge zum Konzept (z. B. «Man kann beide Federwaagen nehmen. Die Wahl der Federwaage hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe Federwaage B genommen, weil ich Angst hatte, dass Federwaage A kaputt geht», weil hier die Auswahl des Messinstruments vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe Federwaage A genommen, weil diese übersichtlicher ist, man kann es besser ablesen.»							
S.: «Ich habe Federwaage A genommen, weil diese vom Messbereich ausreicht.»							
S.: «Ich habe Federwaage A genommen, weil diese genauer ist und vom Messbereich ausreicht.»							
S.: «Ich habe Federwaage A genommen, weil diese in g anzeigt und Federwaage B in kg.»							
S.: «Ich habe Federwaage A genommen, weil diese genauer ist.» Mit Verweis auf die feinere Skala (z. B. Federwaage A zeigt mehr Stellen an).							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe mit beiden Federwaagen gemessen. Zum Vergleichen.»							
S.: «Ich habe Federwaage B genommen, weil diese mehr Belastung aushält. Ich hatte Angst, dass Federwaage A kaputt geht.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

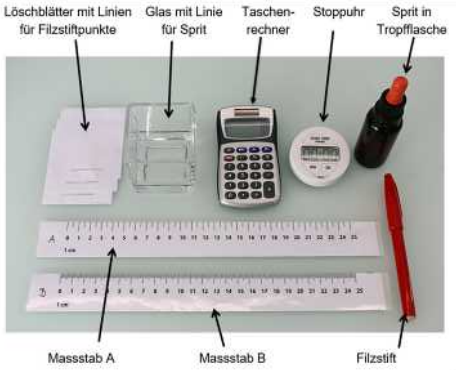
Filzstiftaufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, wie weit Filzstiftpunkte in Sprit in 2 Minuten wandern.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Massstab A: 5 mm-Skala
- Massstab B: 1 mm-Skala
- Filzstiftpunkte wandern in Sprit in 2 Minuten circa 0.6 cm bis 2 cm.



	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Messwiederholung bedeutet bei dieser Aufgabe mehrmals hintereinander mit einem neuen Löschblatt zu messen.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (**mit mehreren Filzstiftpunkten auf einmal**) zu messen, um ein möglichst genaues Ergebnis zu erhalten.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (**z. B. mit 4 min anstelle von 2 min**) zu messen, um ein möglichst genaues Ergebnis zu erhalten.				

Möglichkeit für Rückmeldungen zur Einschätzung:

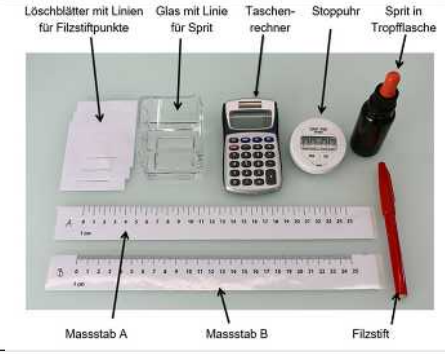
Filzstiftaufgabe – 2a

Die Schülerinnen und Schüler sollen herausfinden, wie weit Filzstiftpunkte in Sprit in 2 Minuten wandern.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Massstab A: 5 mm-Skala
- Massstab B: 1 mm-Skala
- Filzstiftpunkte wandern in Sprit in 2 Minuten circa 0.6 cm bis 2 cm.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Filzstiftaufgabe geht es um das Konzept: Wenn mit einer Menge (z. B. 4 min anstelle von 2 min) gemessen wird, dann erhöht dies die Messgenauigkeit, weil die zurückgelegte Strecke genauer abgelesen werden kann. Da nicht die fachinhaltliche, sondern die fachmethodische Kompetenz überprüft werden soll, ist es in Ordnung, wenn der / die S. dabei annimmt, dass die gemessene Zeit und die zurückgelegte Strecke linear zusammenhängen.

Zu prüfen wäre also bei der Filzstiftaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der gemessenen Zeit und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe mit 4 min gemessen, weil so die zurückgelegte Strecke besser ersichtlich ist und es genauer wird») als auch falsche Bezüge zum Konzept (z. B. «Die gemessene Zeit hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe mit 4 min gemessen, weil ich sowieso ausreichend Zeit für das Experiment zur Verfügung hatte», weil hier die gemessene Zeit vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die 'beste' Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. begründet nicht, ob es bei der Aufgabe besser ist, mit 2 min oder mehr als 2 min (z. B. 4 min) zu messen.							
S.: «Es ist bei dieser Aufgabe besser, mit 2 min zu messen.» S. begründet nicht warum.							
S.: «Es ist bei dieser Aufgabe besser, mit 2 min zu messen, weil es so in der Aufgabe steht.»							
S.: «Es ist bei dieser Aufgabe besser, mit mehr als 2 min (z. B. 4 min) zu messen, weil es so genauer wird; man kann die zurückgelegte Strecke besser messen.»							
S.: «Es ist bei dieser Aufgabe besser, mit 1 min zu messen. So kann man Zeit sparen.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

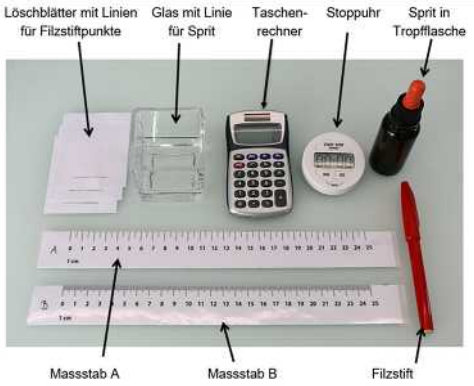
Filzstiftaufgabe – 2b

Die Schülerinnen und Schüler sollen herausfinden, wie weit Filzstiftpunkte in Sprit in 2 Minuten wandern.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Massstab A: 5 mm-Skala
- Massstab B: 1 mm-Skala
- Filzstiftpunkte wandern in Sprit in 2 Minuten circa 0.6 cm bis 2 cm.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Filzstiftaufgabe geht es um das Konzept: Wenn mit einer Menge (mehrere Filzstiftpunkte auf einmal) gemessen wird, dann erhöht dies die Messgenauigkeit. Die Messgenauigkeit wird dabei erhöht, indem man mehrere Anhaltspunkte hat und z. B. einen Mittelwert berechnen kann.

Zu prüfen wäre also bei der Filzstiftaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der Anzahl gemessener Filzstiftpunkte und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe mit mehreren Filzstiftpunkten auf einmal gemessen, weil ich so mehrere Anhaltspunkte habe und einen Mittelwert berechnen kann») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nur mit einem Filzstiftpunkt gemessen. Die Anzahl gemessener Filzstiftpunkte hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe mit mehreren Filzstiftpunkten auf einmal gemessen, weil das Ergebnis schön aussieht», weil hier die Anzahl gemessener Filzstiftpunkte vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die 'beste' Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. begründet nicht, ob es bei dieser Aufgabe besser ist, mit einem Punkt auf einmal oder mehreren Punkten auf einmal zu messen.							
S.: «Es ist bei dieser Aufgabe besser, mit einem Punkt auf einmal zu messen.» S. begründet nicht warum.							
S.: «Es ist bei dieser Aufgabe besser, mit einem Punkt auf einmal zu messen. Weil es steht so in der Aufgabe.»							
S.: «Es ist bei dieser Aufgabe besser, mit einem Punkt auf einmal zu messen. Man weiss nicht, ob es einen Einfluss auf das Ergebnis hat, wenn man mit mehreren Punkten auf einmal misst.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Es ist bei dieser Aufgabe besser, mit mehreren Punkten auf einmal zu messen. So hat man mehrere Anhaltspunkte auf einmal und braucht nicht so viele Löschpapiere.»							
S.: «Es ist bei dieser Aufgabe besser, mit mehreren Punkten auf einmal zu messen. So hat man mehrere Anhaltspunkte und kann sehen, ob die Punkte an allen Stellen auf dem Löschpapier gleich weit wandern oder z. B. am Rand des Papiers weiter wandern.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S: «Es ist bei dieser Aufgabe besser, mit unterschiedlich vielen Punkten auf einmal zu messen (z. B. einmal mit nur einem Punkt, einmal mit mehreren Punkten auf einmal, etc.). So kann man untersuchen, ob die Anzahl Punkte einen Einfluss hat.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

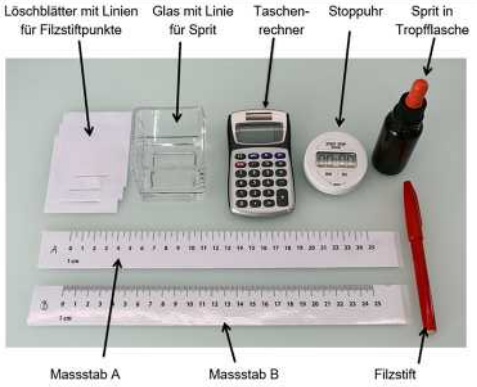
Filzstiftaufgabe – 3

Die Schülerinnen und Schüler sollen herausfinden, wie weit Filzstiftpunkte in Sprit in 2 Minuten wandern.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Massstab A: 5 mm-Skala
- Massstab B: 1 mm-Skala
- Filzstiftpunkte wandern in Sprit in 2 Minuten circa 0.6 cm bis 2 cm.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Filzstiftaufgabe geht es um das Konzept: Wenn das genauere Messinstrument (hier Massstab B) verwendet wird, dann erhöht dies die Messgenauigkeit.

Zu prüfen wäre also bei der Filzstiftaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen dem gewählten Messinstrument und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Massstab B genommen, weil dieser genauer ist») als auch falsche Bezüge zum Konzept (z. B. «Man kann beide Massstäbe nehmen. Die Wahl des Massstabs hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe Massstab B genommen, weil ich Massstab A nicht sofort am Arbeitsplatz gefunden habe», weil hier die Auswahl des Messinstruments vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe Massstab B genommen, weil diese genauer ist.» Mit Verweis auf die feinere Skala.							

Möglichkeit für Rückmeldungen zur Einschätzung:

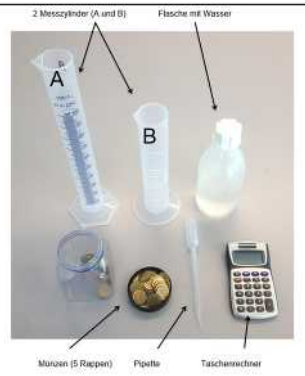
Münzenaufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, wie gross die Verdrängung einer 5 Rappen-Münze ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Messzylinder A: Fassungsvermögen 100 ml, Skala auf 1 ml genau.
- Messzylinder B: Fassungsvermögen 100 ml, Skala auf 5 ml genau.
- Eine 5 Rappen-Münze hat eine Verdrängung von circa 0.1 ml bis 0.4 ml.



	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Messwiederholung bedeutet bei dieser Aufgabe, mehrmals hintereinander mit einer bestimmten Anzahl Münzen zu messen.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (z. B. 10 Münzen auf einmal) zu messen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Die Verdrängung einer einzelnen Münze kann mit den gegebenen Messinstrumenten kaum ermittelt werden und wird sehr ungenau. Wenn mit einer Menge gemessen wird (z. B. 10 Münzen auf einmal), wird das Ergebnis genauer.				

Möglichkeit für Rückmeldungen zur Einschätzung:

Münzenaufgabe – 2

Die Schülerinnen und Schüler sollen herausfinden, wie gross die Verdrängung einer 5 Rappen-Münze ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Messzylinder A: Fassungsvermögen 100 ml, Skala auf 1 ml genau.
- Messzylinder B: Fassungsvermögen 100 ml, Skala auf 5 ml genau.
- Eine 5 Rappen-Münze hat eine Verdrängung von circa 0.1 ml bis 0.4 ml.

Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Münzenaufgabe geht es um das Konzept: Wenn mit einer Menge gemessen wird (z. B. 10 Münzen auf einmal), dann erhöht dies die Messgenauigkeit. Dabei gilt, dass je grösser die Menge (z. B. 20 Münzen auf einmal), desto grösser die Messgenauigkeit. Es kann mit einer Menge gemessen werden, weil das Experiment so genauer durchführbar ist (z. B. mit 10 Münzen messen, damit man es auf dem Messzylinder genauer ablesen kann) oder weil es tatsächlich genauer wird (z. B. mit 20 Münzen messen, damit man minimale Unterschiede bei den Münzen berücksichtigen kann). Diese Gründe für das Messen mit einer Menge widerspiegeln dabei unterschiedliche Niveaus der S.

Zu prüfen wäre also bei der Münzenaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der Anzahl gemessener Münzen und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe 20 Münzen auf einmal genommen, damit die Messung genauer wird» oder «Ich habe 10 Münzen genommen, damit ich es auf der Skala genauer ablesen kann») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nicht auf die Anzahl Münzen geachtet, weil diese keinen Einfluss auf die Genauigkeit hat»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe nur eine Münze genommen, weil ich nicht abzählen wollte», weil hier die Auswahl der Anzahl der Münzen vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. sagt, dass es besser ist, mit mehr als einer Münze auf einmal zu messen. Ohne weitere Begründung.							
S.: «Es ist bei dieser Aufgabe besser, mit mehr als einer Münze auf einmal zu messen (z. B. 5 Münzen auf einmal). Bei einer Münze sieht man noch keine Verdrängung, mit mehreren Münzen kann man es beim Messzylinder besser ablesen.»							
S.: «Es ist besser, mit mehr als einer Münze zu messen. Denn wenn man die Verdrängung für eine Münze misst, wird es sehr ungenau.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

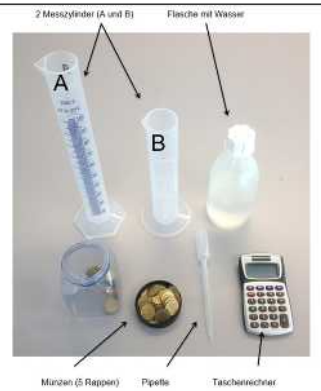
Münzenaufgabe – 3

Die Schülerinnen und Schüler sollen herausfinden, wie gross die Verdrängung einer 5 Rappen-Münze ist.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Messzylinder A: Fassungsvermögen 100 ml, Skala auf 1 ml genau.
- Messzylinder B: Fassungsvermögen 100 ml, Skala auf 5 ml genau.
- Eine 5 Rappen-Münze hat eine Verdrängung von circa 0.1 ml bis 0.4 ml.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Münzenaufgabe geht es um das Konzept: Wenn das genauere Messinstrument (hier Messzylinder A) verwendet wird, dann erhöht dies die Messgenauigkeit.

Zu prüfen wäre also bei der Münzenaufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen dem gewählten Messinstrument und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Messzylinder A genommen, weil dieser genauer ist») als auch falsche Bezüge zum Konzept (z. B. «Man kann beide Messzylinder nehmen. Die Wahl des Messzylinders hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe Messzylinder A genommen, weil ich die Skala bei Messzylinder B nicht verstanden habe», weil hier die Auswahl des Messinstruments vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. verwendet beide Messinstrumente für die Messung. Ohne weitere Begründung warum dies so gemacht wurde.							
S.: «Ich habe Messzylinder A genommen, weil dieser genauer ist.» Ohne weitere Begründung warum dies so ist.							
S.: «Ich habe Messzylinder A genommen, weil dieser genauer ist.» Mit Verweis auf die feinere Skala (z. B. Messzylinder A zeigt auf 1 ml genau an, etc.).							

Möglichkeit für Rückmeldungen zur Einschätzung:

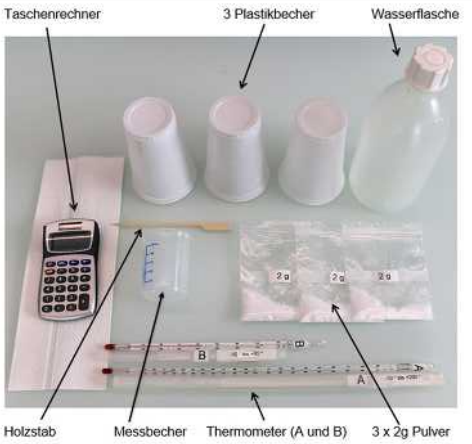
Pulveraufgabe – 1

Die Schülerinnen und Schüler sollen herausfinden, wie sich die Temperatur von 50 ml Wasser verändert, wenn sie ein Pulver hineingeben.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Thermometer A: Messbereich - 10 °C bis + 200 °C; Skala auf 1 °C genau.
- Thermometer B: Messbereich -10 °C bis + 60 °C; Skala auf 1 °C genau. Abstände zwischen Skalenstrichen grösser, somit auch die Möglichkeit auf circa 0.5 °C genau abzulesen.
- Die Temperatur von 50 ml Wasser verändert sich um circa 1.5 °C bis 3.5 °C, wenn darin das Pulver aufgelöst wird.



	Nicht naheliegend	Eher nicht naheliegend	Eher naheliegend	Naheliegend
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe Messwiederholungen durchzuführen, um ein möglichst genaues Ergebnis zu erhalten.				
Schätzen Sie ein, wie naheliegend es für Jugendliche ist, bei dieser Aufgabe mit einer Menge (z. B. 2 Tütchen Pulver auf einmal) zu messen, um ein möglichst genaues Ergebnis zu erhalten. *Hinweis:* Mit 2 Tütchen Pulver auf einmal ist der Temperaturunterschied deutlicher und kann somit genauer abgelesen werden.				

Möglichkeit für Rückmeldungen zur Einschätzung:

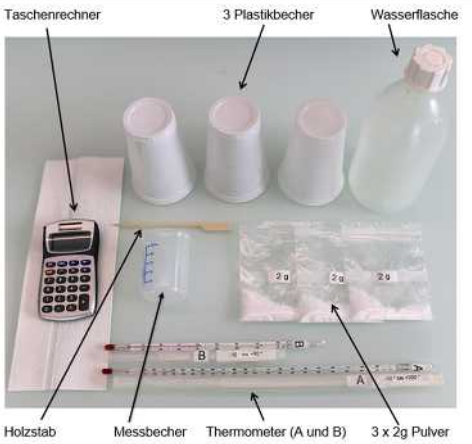
Pulveraufgabe – 2

Die Schülerinnen und Schüler sollen herausfinden, wie sich die Temperatur von 50 ml Wasser verändert, wenn sie ein Pulver hineingeben.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Thermometer A: Messbereich - 10 °C bis + 200 °C; Skala auf 1 °C genau.
- Thermometer B: Messbereich -10 °C bis + 60 °C; Skala auf 1 °C genau. Abstände zwischen Skalenstrichen grösser, somit auch die Möglichkeit auf circa 0.5 °C genau abzulesen.
- Die Temperatur von 50 ml Wasser verändert sich um circa 1.5 °C bis 3.5 °C, wenn darin das Pulver aufgelöst wird.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Pulveraufgabe geht es um das Konzept: Wenn mit einer Menge (z. B. 2 Tütchen auf einmal) gemessen wird, dann erhöht dies die Messgenauigkeit. Die Messgenauigkeit kann dadurch erhöht werden, dass der Temperaturunterschied eindeutiger ausfällt und dadurch genauer abgelesen werden kann. Da nicht die fachinhaltliche, sondern die fachmethodische Kompetenz überprüft werden soll, ist es in Ordnung, wenn der / die S. dabei annimmt, dass der Temperaturunterschied und die Menge an Pulver linear zusammenhängen.

Zu prüfen wäre also bei der Pulveraufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen der Menge an Pulver und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe mit mehreren Tütchen auf einmal gemessen, weil ich so den Temperaturunterschied deutlicher ablesen kann und es somit genauer wird») als auch falsche Bezüge zum Konzept (z. B. «Ich habe nur mit einem Tütchen auf einmal gemessen. Die Menge an Pulver hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe mit 3 Tütchen auf einmal gemessen, weil ich 3 Tütchen zur Verfügung hatte», weil hier die gemessene Menge an Pulver vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
Keine Aussage, ob es besser ist, mit 1 Tütchen oder mit 2 Tütchen auf einmal zu messen.							
S.: «Es ist besser, mit 1 Tütchen auf einmal zu messen.» Ohne Begründung warum das so ist.							
S.: «Es ist besser, mit 1 Tütchen auf einmal zu messen. Weil es so in der Aufgabe steht.»							
S.: «Es ist besser, mit 1 Tütchen auf einmal zu messen. Weil es mit 2 Tütchen ungenau werden könnte.»							
S: «Es ist besser, mit 1 Tütchen auf einmal zu messen. Man hat ja nur 3 Tütchen und so kann man die Messung dreimal wiederholen.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. «Ich habe nur mit 1 Tütchen auf einmal gemessen, weil ich dachte, man darf nur 1 Tütchen verwenden.»							
S.: «Es ist besser, zuerst mit 1 Tütchen zu messen, dann ein weiteres Tütchen ins Wasser zu geben. So kann man vergleichen.»							
S: «Es ist besser, mit 2 Tütchen auf einmal zu messen. Weil sich die Temperatur mit 1 Tütchen kaum verändert, mit 2 Tütchen kann man es beim Thermometer besser ablesen.»							
S.: «Ich habe mit mehreren Tütchen gemessen. Ich wollte sehen, wie es sich verändert, weil es mich interessiert hat / Neugierde.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

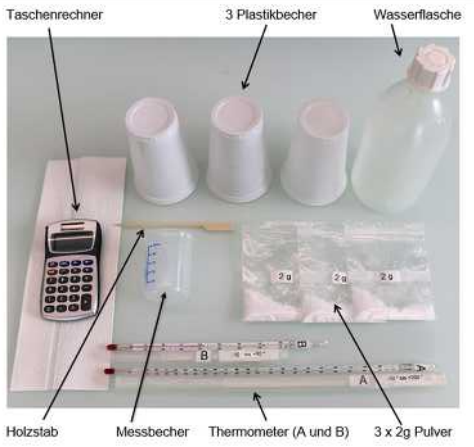
Pulveraufgabe – 3

Die Schülerinnen und Schüler sollen herausfinden, wie sich die Temperatur von 50 ml Wasser verändert, wenn sie ein Pulver hineingeben.

Dazu erhalten sie das abgebildete Material.

Hinweise für Expertinnen und Experten

- Thermometer A: Messbereich - 10 °C bis + 200 °C; Skala auf 1 °C genau.
- Thermometer B: Messbereich -10 °C bis + 60 °C; Skala auf 1 °C genau. Abstände zwischen Skalenstrichen grösser, somit auch die Möglichkeit auf circa 0.5 °C genau abzulesen.
- Die Temperatur von 50 ml Wasser verändert sich um circa 1.5 °C bis 3.5 °C, wenn darin das Pulver aufgelöst wird.



Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgabe über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei der folgenden Einschätzung zur Pulveraufgabe geht es um das Konzept: Wenn das genauere Messinstrument (hier Thermometer B) verwendet wird, dann erhöht dies die Messgenauigkeit.

Zu prüfen wäre also bei der Pulveraufgabe, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgabe über den **Zusammenhang zwischen dem gewählten Messinstrument und der Genauigkeit der Messung** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Thermometer B genommen, weil dieses genauer ist») als auch falsche Bezüge zum Konzept (z. B. «Man kann beide Thermometer nehmen. Die Wahl des Thermometers hat keinen Einfluss auf die Messgenauigkeit»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe Thermometer A genommen, weil Thermometer B nicht richtig anzeigte», weil hier die Auswahl des Messinstruments vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S. verwendet Thermometer A, weil dieses besser zu bedienen ist (z. B. weil es länger ist).							
S. verwendet Thermometer A, weil Thermometer B nicht richtig funktionierte.							
S.: «Ich habe Thermometer A genommen, weil ich nicht wusste wie stark sich die Temperatur verändert. Darum habe ich zur Sicherheit den grösseren Messbereich genommen.»							
S: «Ich habe Thermometer A genommen. Dieses sieht genauer aus, weil es grösser ist.»							
S.: «Ich habe beide Thermometer genommen. So konnte ich zwei Messungen gleichzeitig durchführen.»							
S: «Ich habe beide Thermometer genommen, zum Vergleichen.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe Thermometer B genommen, weil dieses vom Messbereich ausreicht und genauer ist.»							
S.: «Ich habe Thermometer B genommen, weil dieses besser zu bedienen ist». Z. B. weil es kürzer ist und so der Becher nicht umkippt.							
S.: «Ich habe Thermometer B genommen, weil dieses genauer ist. Bei Thermometer B sind die Abstände auf der Skala grösser, man kann so auch auf 0.5 °C genau ablesen.»							

Möglichkeit für Rückmeldungen zur Einschätzung:

Allgemein Messwiederholung

Schätzen Sie anhand der kategorisierten Schülerinnen- und Schüleraussagen ein, ob es Hinweise gibt, dass der Schüler bzw. die Schülerin (S.) bei der Bearbeitung der Aufgaben des Problemtyps «Messen mit vorgegebenen Instrumenten» über das intendierte Konzept nachgedacht hat. Entscheidend ist dabei nur die Frage, **ob** ein Bezug zu diesem Konzept hergestellt wurde und **nicht**, ob das Konzept richtig angewendet wurde.

Intendiertes Konzept

Bei den Aufgaben des Problemtyps «Messen mit vorgegebenen Instrumenten» (Ahorn-, Bohnen-, Faden-, Filzstift-, Münzen- und Pulveraufgabe) geht es auch um das Konzept: Wenn Messungen wiederholt werden, dann erhöht dies die Messgenauigkeit. Die Messgenauigkeit kann erhöht werden, indem z. B. ein Mittelwert als Resultat berechnet wird; ein Intervall angegeben wird (also gesagt wird in welchem Bereich die zu messende Grösse liegt); ein mittlerer Wert ausgewählt wird (Median) oder nochmals nachgemessen wird, als eine Art Bestätigung des vorherigen Ergebnisses. Diese Gründe für das Durchführen von Messwiederholungen befinden sich auf unterschiedlichen Niveaus: Während das Berechnen eines Mittelwerts eher einem hohen Niveau entspricht, entspricht das Wiederholen der Messung zur Bestätigung eher einem niedrigen Niveau.

Zu prüfen wäre nun, ob es bei den kategorisierten S.- Aussagen Hinweise gibt, dass bei der Bearbeitung der Aufgaben des Problemtyps «Messen mit vorgegebenen Instrumenten» über den **Zusammenhang zwischen dem Wiederholen von Messungen und der Messgenauigkeit** nachgedacht wurde. Dazu gehören sowohl richtige (z. B. «Ich habe Messungen wiederholt, um ein genaueres Ergebnis zu erhalten» oder «Ich habe Messungen wiederholt, um zu sehen in welchem Bereich die Ergebnisse sind») als auch falsche Bezüge zum Konzept (z. B. «Es spielt keine Rolle ob man einmal oder mehrmals misst, wenn man das Experiment genau durchgeführt hat»). Ein Hinweis auf die Nutzung eines nicht intendierten Konzepts wäre beispielsweise: «Ich habe mehrmals gemessen, weil ich ausreichend Zeit hatte», weil hier das Durchführen von Messwiederholungen vermutlich nicht mit der Messgenauigkeit verknüpft wurde.

Hinweis: Falls es bei der kategorisierten S.- Aussage Hinweise für mehrere Konzepte gibt, dann bitte ich Sie, im Rating immer die ‘beste’ Einschätzung vorzunehmen (z. B. wenn es Hinweise auf intendierte und nicht intendierte Konzepte gibt, dann bitte ich Sie, das intendierte Konzept einzuschätzen).

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
Mehrmals messen ist besser. Ohne Begründung warum das so ist.							
S.: «Mehrmals messen ist besser. Weil man so sein Ergebnis bestätigen kann.»							
S.: «Mehrmals messen ist besser. Weil sich so Ungenauigkeiten ausgleichen (z. B. bei Ahornaufgabe: Samen fallen nicht immer gleich; bei Fadenaufgabe: Der Faden ist nicht an allen Stellen gleich; Bohnenaufgabe: Bohnen sind unterschiedlich gross; etc.).»							
S.: «Mehrmals messen ist besser. Weil man so sehen kann, in welchem Bereich die Zahlen sind.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Mehrmals messen ist besser. So wird das Ergebnis genauer.» Ohne Begründung warum das so ist.							
S.: «Mehrmals messen ist besser. Weil man so ein genaueres Ergebnis kriegt und einen Mittelwert berechnen kann.»							
S.: «Mehrmals messen ist besser. Man hat ausreichend Zeit dafür.»							
S.: «Mehrmals messen ist besser. Bei der ersten Messung hatte ich noch Schwierigkeiten, nachher wusste ich wie es geht.»							
S.: «Bei dieser Aufgabe ist mehrmals messen besser.» Mit unlogischer Begründung (z. B. Ahornaufgabe: Je mehr man misst, desto schneller fliegt der Samen).							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Ich habe mehrmals gemessen, weil ich einen Zusammenhang untersuchen wollte.» Z. B. bei Filzstiftaufgabe: Ich wollte schauen, ob die Punkte am Rand und in der Mitte des Löschpapiers gleich weit wandern; bei Ahornaufgabe: Ich wollte die beiden Samen miteinander vergleichen und schauen, ob die Zeit mit der Samengrösse zusammenhängt; etc.							
S.: «Man kriegt das genauste Ergebnis, wenn man es nur einmal, dafür aber richtig macht.»							
S.: «Einmal messen genügt, wenn das Experiment genau durchgeführt wurde. Es würde sowieso in etwa das gleiche Ergebnis herauskommen.»							

	Es kann nicht beurteilt werden, über welche Konzepte bei der Bearbeitung der Aufgabe nachgedacht wurde	Es wurde primär über ein Konzept nachgedacht, das **nicht** intendiert ist	Es gibt Hinweise, dass über das intendierte Konzept nachgedacht wurde.				
			Falscher Bezug zum Konzept	Richtiger Bezug zum Konzept, niedriges Niveau	Richtiger Bezug zum Konzept, eher niedriges Niveau	Richtiger Bezug zum Konzept, eher hohes Niveau	Richtiger Bezug zum Konzept, hohes Niveau
S.: «Einmal messen genügt, weil man zu wenig Zeit zum mehrmals messen hat.»							
S.: «Einmal messen genügt. Die Aufgabe verlangt nicht, dass man mehrmals misst.»							
S.: «Ich habe nur einmal gemessen, weil ich dachte, dass man nur einmal messen darf.» (z. B. bei der Filzstift- oder Pulveraufgabe: Weil gedacht wurde, dass nur 1 Löschpapier bzw. 1 Tütchen verwendet werden darf).							

Möglichkeit für Rückmeldungen zur Einschätzung:

Vielen Dank, dass Sie sich die Zeit genommen haben, am Expertenrating teilzunehmen!

☺

Bisher erschienene Bände der Reihe *„Studien zum Physik- und Chemielernen“*

ISSN 1614-8967 (vormals *Studien zum Physiklernen* ISSN 1435-5280)

1 Helmut Fischler, Jochen Peuckert (Hrsg.): Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie
ISBN 978-3-89722-256-4 40.50 EUR

2 Anja Schoster: Bedeutungsentwicklungsprozesse beim Lösen algorithmischer Physikaufgaben. *Eine Fallstudie zu Lernprozessen von Schülern im Physiknachhilfeunterricht während der Bearbeitung algorithmischer Physikaufgaben*
ISBN 978-3-89722-045-4 40.50 EUR

3 Claudia von Aufschnaiter: Bedeutungsentwicklungen, Interaktionen und situatives Erleben beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-89722-143-7 40.50 EUR

4 Susanne Haeberlen: Lernprozesse im Unterricht mit Wasserstromkreisen. *Eine Fallstudie in der Sekundarstufe I*
ISBN 978-3-89722-172-7 40.50 EUR

5 Kerstin Haller: Über den Zusammenhang von Handlungen und Zielen. *Eine empirische Untersuchung zu Lernprozessen im physikalischen Praktikum*
ISBN 978-3-89722-242-7 40.50 EUR

6 Michaela Horstendahl: Motivationale Orientierungen im Physikunterricht
ISBN 978-3-89722-227-4 50.00 EUR

7 Stefan Deylitz: Lernergebnisse in der Quanten-Atomphysik. *Evaluation des Bremer Unterrichtskonzepts*
ISBN 978-3-89722-291-5 40.50 EUR

8 Lorenz Hucke: Handlungsregulation und Wissenserwerb in traditionellen und computergestützten Experimenten des physikalischen Praktikums
ISBN 978-3-89722-316-5 50.00 EUR

9 Heike Theyßen: Ein Physikpraktikum für Studierende der Medizin. *Darstellung der Entwicklung und Evaluation eines adressatenspezifischen Praktikums nach dem Modell der Didaktischen Rekonstruktion*
ISBN 978-3-89722-334-9 40.50 EUR

10 Annette Schick: Der Einfluß von Interesse und anderen selbstbezogenen Kognitionen auf Handlungen im Physikunterricht. *Fallstudien zu Interessenhandlungen im Physikunterricht*
ISBN 978-3-89722-380-6 40.50 EUR

11 Roland Berger: Moderne bildgebende Verfahren der medizinischen Diagnostik. *Ein Weg zu interessanterem Physikunterricht*
ISBN 978-3-89722-445-2 40.50 EUR

12 Johannes Werner: Vom Licht zum Atom. *Ein Unterrichtskonzept zur Quantenphysik unter Nutzung des Zeigermodells*
ISBN 978-3-89722-471-1 40.50 EUR

13 Florian Sander: Verbindung von Theorie und Experiment im physikalischen Praktikum. *Eine empirische Untersuchung zum handlungsbezogenen Vorverständnis und dem Einsatz grafikorientierter Modellbildung im Praktikum*
ISBN 978-3-89722-482-7 40.50 EUR

14 Jörn Gerdes: Der Begriff der physikalischen Kompetenz. *Zur Validierung eines Konstruktes*
ISBN 978-3-89722-510-7 40.50 EUR

15 Malte Meyer-Arndt: Interaktionen im Physikpraktikum zwischen Studierenden und Betreuern. *Feldstudie zu Bedeutungsentwicklungsprozessen im physikalischen Praktikum*
ISBN 978-3-89722-541-1 40.50 EUR

16 Dietmar Höttecke: Die Natur der Naturwissenschaften historisch verstehen. *Fachdidaktische und wissenschaftshistorische Untersuchungen*
ISBN 978-3-89722-607-4 40.50 EUR

17 Gil Gabriel Mavanga: Entwicklung und Evaluation eines experimentell- und phänomenorientierten Optikcurriculums. *Untersuchung zu Schülervorstellungen in der Sekundarstufe I in Mosambik und Deutschland*
ISBN 978-3-89722-721-7 40.50 EUR

18 Meike Ute Zastrow: Interaktive Experimentieranleitungen. *Entwicklung und Evaluation eines Konzeptes zur Vorbereitung auf das Experimentieren mit Messgeräten im Physikalischen Praktikum*
ISBN 978-3-89722-802-3 40.50 EUR

19 Gunnar Friege: Wissen und Problemlösen. *Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*
ISBN 978-3-89722-809-2 40.50 EUR

20 Erich Starauschek: Physikunterricht nach dem Karlsruher Physikkurs. *Ergebnisse einer Evaluationsstudie*
ISBN 978-3-89722-823-8 40.50 EUR

21 Roland Paatz: Charakteristika analogiebasierten Denkens. *Vergleich von Lernprozessen in Basis- und Zielbereich*
ISBN 978-3-89722-944-0 40.50 EUR

22 Silke Mikelskis-Seifert: Die Entwicklung von Metakonzepten zur Teilchenvorstellung bei Schülern. *Untersuchung eines Unterrichts über Modelle mithilfe eines Systems multipler Repräsentationsebenen*
ISBN 978-3-8325-0013-9 40.50 EUR

23 Brunhild Landwehr: Distanzen von Lehrkräften und Studierenden des Sachunterrichts zur Physik. *Eine qualitativ-empirische Studie zu den Ursachen*
ISBN 978-3-8325-0044-3 40.50 EUR

24 Lydia Murmann: Physiklernen zu Licht, Schatten und Sehen. *Eine phänomenografische Untersuchung in der Primarstufe*
ISBN 978-3-8325-0060-3 40.50 EUR

25 Thorsten Bell: Strukturprinzipien der Selbstregulation. *Komplexe Systeme, Elementarisierungen und Lernprozessstudien für den Unterricht der Sekundarstufe II*
ISBN 978-3-8325-0134-1 40.50 EUR

26 Rainer Müller: Quantenphysik in der Schule
ISBN 978-3-8325-0186-0 40.50 EUR

27 Jutta Roth: Bedeutungsentwicklungsprozesse von Physikerinnen und Physikern in den Dimensionen Komplexität, Zeit und Inhalt
ISBN 978-3-8325-0183-9 40.50 EUR

28 Andreas Saniter: Spezifika der Verhaltensmuster fortgeschrittener Studierender der Physik
ISBN 978-3-8325-0292-8 40.50 EUR

29 Thomas Weber: Kumulatives Lernen im Physikunterricht. *Eine vergleichende Untersuchung in Unterrichtsgängen zur geometrischen Optik*
ISBN 978-3-8325-0316-1 40.50 EUR

30 Markus Rehm: Über die Chancen und Grenzen moralischer Erziehung im naturwissenschaftlichen Unterricht
ISBN 978-3-8325-0368-0 40.50 EUR

31 Marion Budde: Lernwirkungen in der Quanten-Atom-Physik. *Fallstudien über Resonanzen zwischen Lernangeboten und SchülerInnen-Vorstellungen*
ISBN 978-3-8325-0483-0 40.50 EUR

32 Thomas Reyer: Oberflächenmerkmale und Tiefenstrukturen im Unterricht. *Exemplarische Analysen im Physikunterricht der gymnasialen Sekundarstufe*
ISBN 978-3-8325-0488-5 40.50 EUR

33 Christoph Thomas Müller: Subjektive Theorien und handlungsleitende Kognitionen von Lehrern als Determinanten schulischer Lehr-Lern-Prozesse im Physikunterricht
ISBN 978-3-8325-0543-1 40.50 EUR

34 Gabriela Jonas-Ahrend: Physiklehrervorstellungen zum Experiment im Physikunterricht
ISBN 978-3-8325-0576-9 40.50 EUR

35 Dimitrios Stavrou: Das Zusammenspiel von Zufall und Gesetzmäßigkeiten in der nichtlinearen Dynamik. *Didaktische Analyse und Lernprozesse*
ISBN 978-3-8325-0609-4 40.50 EUR

36 Katrin Engeln: Schülerlabors: authentische, aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken
ISBN 978-3-8325-0689-6 40.50 EUR

37 Susann Hartmann: Erklärungsvielfalt
ISBN 978-3-8325-0730-5 40.50 EUR

38 Knut Neumann: Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker
ISBN 978-3-8325-0762-6 40.50 EUR

39 Michael Späth: Kontextbedingungen für Physikunterricht an der Hauptschule. *Möglichkeiten und Ansatzpunkte für einen fachübergreifenden, handlungsorientierten und berufsorientierten Unterricht*
ISBN 978-3-8325-0827-2 40.50 EUR

40 Jörg Hirsch: Interesse, Handlungen und situatives Erleben von Schülerinnen und Schülern beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-8325-0875-3 40.50 EUR

41 Monika Hüther: Evaluation einer hypermedialen Lernumgebung zum Thema Gasgesetze. *Eine Studie im Rahmen des Physikpraktikums für Studierende der Medizin*
ISBN 978-3-8325-0911-8 40.50 EUR

42 Maike Tesch: Das Experiment im Physikunterricht. *Didaktische Konzepte und Ergebnisse einer Videostudie*
ISBN 978-3-8325-0975-0 40.50 EUR

43 Nina Nicolai: Skriptgeleitete Eltern-Kind-Interaktion bei Chemiehausaufgaben. *Eine Evaluationsstudie im Themenbereich Säure-Base*
ISBN 978-3-8325-1013-8 40.50 EUR

44 Antje Leisner: Entwicklung von Modellkompetenz im Physikunterricht
ISBN 978-3-8325-1020-6 40.50 EUR

45 Stefan Rumann: Evaluation einer Interventionsstudie zur Säure-Base-Thematik
ISBN 978-3-8325-1027-5 40.50 EUR

46 Thomas Wilhelm: Konzeption und Evaluation eines Kinematik/Dynamik-Lehrgangs zur Veränderung von Schülervorstellungen mit Hilfe dynamisch ikonischer Repräsentationen und graphischer Modellbildung – mit CD-ROM
ISBN 978-3-8325-1046-6 45.50 EUR

47 Andrea Maier-Richter: Computerunterstütztes Lernen mit Lösungsbeispielen in der Chemie. *Eine Evaluationsstudie im Themenbereich Löslichkeit*
ISBN 978-3-8325-1046-6 40.50 EUR

48 Jochen Peuckert: Stabilität und Ausprägung kognitiver Strukturen zum Atombegriff
ISBN 978-3-8325-1104-3 40.50 EUR

49 Maik Walpuski: Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback
ISBN 978-3-8325-1184-5 40.50 EUR

50 Helmut Fischler, Christiane S. Reiners (Hrsg.): Die Teilchenstruktur der Materie im Physik- und Chemieunterricht
ISBN 978-3-8325-1225-5 34.90 EUR

51 Claudia Eysel: Interdisziplinäres Lehren und Lernen in der Lehrerbildung. *Eine empirische Studie zum Kompetenzerwerb in einer komplexen Lernumgebung*
ISBN 978-3-8325-1238-5 40.50 EUR

52 Johannes Günther: Lehrerfortbildung über die Natur der Naturwissenschaften. *Studien über das Wissenschaftsverständnis von Grundschullehrkräften*
ISBN 978-3-8325-1287-3 40.50 EUR

53 Christoph Neugebauer: Lernen mit Simulationen und der Einfluss auf das Problemlösen in der Physik
ISBN 978-3-8325-1300-9 40.50 EUR

54 Andreas Schnirch: Gendergerechte Interessen- und Motivationsförderung im Kontext naturwissenschaftlicher Grundbildung. *Konzeption, Entwicklung und Evaluation einer multimedial unterstützten Lernumgebung*
ISBN 978-3-8325-1334-4 40.50 EUR

55 Hilde Köster: Freies Explorieren und Experimentieren. *Eine Untersuchung zur selbstbestimmten Gewinnung von Erfahrungen mit physikalischen Phänomenen im Sachunterricht*
ISBN 978-3-8325-1348-1 40.50 EUR

56 Eva Heran-Dörr: Entwicklung und Evaluation einer Lehrerfortbildung zur Förderung der physikdidaktischen Kompetenz von Sachunterrichtslehrkräften
ISBN 978-3-8325-1377-1 40.50 EUR

57 Agnes Szabone Varnai: Unterstützung des Problemlösens in Physik durch den Einsatz von Simulationen und die Vorgabe eines strukturierten Kooperationsformats
ISBN 978-3-8325-1403-7 40.50 EUR

58 Johannes Rethfeld: Aufgabenbasierte Lernprozesse in selbstorganisationsoffenem Unterricht der Sekundarstufe I zum Themengebiet ELEKTROSTATIK. *Eine Feldstudie in vier 10. Klassen zu einer kartenbasierten Lernumgebung mit Aufgaben aus der Elektrostatik*
ISBN 978-3-8325-1416-7 40.50 EUR

59 Christian Henke: Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. *Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven*
ISBN 978-3-8325-1515-7 40.50 EUR

60 Lutz Kasper: Diskursiv-narrative Elemente für den Physikunterricht. *Entwicklung und Evaluation einer multimedialen Lernumgebung zum Erdmagnetismus*
ISBN 978-3-8325-1537-9 40.50 EUR

61 Thorid Rabe: Textgestaltung und Aufforderung zu Selbsterklärungen beim Physiklernen mit Multimedia
ISBN 978-3-8325-1539-3 40.50 EUR

62 Ina Glemnitz: Vertikale Vernetzung im Chemieunterricht. *Ein Vergleich von traditionellem Unterricht mit Unterricht nach Chemie im Kontext*
ISBN 978-3-8325-1628-4 40.50 EUR

63 Erik Einhaus: Schülerkompetenzen im Bereich Wärmelehre. *Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*
ISBN 978-3-8325-1630-7 40.50 EUR

64 Jasmin Neuroth: Concept Mapping als Lernstrategie. *Eine Interventionsstudie zum Chemielernen aus Texten*
ISBN 978-3-8325-1659-8 40.50 EUR

65 Hans Gerd Hegeler-Burkhart: Zur Kommunikation von Hauptschülerinnen und Hauptschülern in einem handlungsorientierten und fächerübergreifenden Unterricht mit physikalischen und technischen Inhalten
ISBN 978-3-8325-1667-3 40.50 EUR

66 Karsten Rincke: Sprachentwicklung und Fachlernen im Mechanikunterricht. *Sprache und Kommunikation bei der Einführung in den Kraftbegriff*
ISBN 978-3-8325-1699-4 40.50 EUR

67 Nina Strehle: Das Ion im Chemieunterricht. *Alternative Schülervorstellungen und curriculare Konsequenzen*
ISBN 978-3-8325-1710-6 40.50 EUR

68 Martin Hopf: Problemorientierte Schülerexperimente
ISBN 978-3-8325-1711-3 40.50 EUR

69 Anne Beerenwinkel: Fostering conceptual change in chemistry classes using expository texts
ISBN 978-3-8325-1721-2 40.50 EUR

70 Roland Berger: Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II. *Eine empirische Untersuchung auf der Grundlage der Selbstbestimmungstheorie der Motivation*
ISBN 978-3-8325-1732-8 40.50 EUR

71 Giuseppe Colicchia: Physikunterricht im Kontext von Medizin und Biologie. *Entwicklung und Erprobung von Unterrichtseinheiten*
ISBN 978-3-8325-1746-5 40.50 EUR

72 Sandra Winheller: Geschlechtsspezifische Auswirkungen der Lehrer-Schüler-Interaktion im Chemieanfangsunterricht
ISBN 978-3-8325-1757-1 40.50 EUR

73 Isabel Wahser: Training von naturwissenschaftlichen Arbeitsweisen zur Unterstützung experimenteller Kleingruppenarbeit im Fach Chemie
ISBN 978-3-8325-1815-8 40.50 EUR

74 Claus Brell: Lernmedien und Lernerfolg - reale und virtuelle Materialien im Physikunterricht. *Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulation und IBE*
ISBN 978-3-8325-1829-5 40.50 EUR

75 Rainer Wackermann: Überprüfung der Wirksamkeit eines Basismodell-Trainings für Physiklehrer
ISBN 978-3-8325-1882-0 40.50 EUR

76 Oliver Tepner: Effektivität von Aufgaben im Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-1919-3 40.50 EUR

77 Claudia Geyer: Museums- und Science-Center-Besuche im naturwissenschaftlichen Unterricht aus einer motivationalen Perspektive. *Die Sicht von Lehrkräften und Schülerinnen und Schülern*
ISBN 978-3-8325-1922-3 40.50 EUR

78 Tobias Leonhard: Professionalisierung in der Lehrerbildung. *Eine explorative Studie zur Entwicklung professioneller Kompetenzen in der Lehrererstausbildung*
ISBN 978-3-8325-1924-7 40.50 EUR

79 Alexander Kauertz: Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben
ISBN 978-3-8325-1925-4 40.50 EUR

80 Regina Hübinger: Schüler auf Weltreise. *Entwicklung und Evaluation von Lehr-/ Lernmaterialien zur Förderung experimentell-naturwissenschaftlicher Kompetenzen für die Jahrgangsstufen 5 und 6*
ISBN 978-3-8325-1932-2 40.50 EUR

81 Christine Waltner: Physik lernen im Deutschen Museum
ISBN 978-3-8325-1933-9 40.50 EUR

82 Torsten Fischer: Handlungsmuster von Physiklehrkräften beim Einsatz neuer Medien. *Fallstudien zur Unterrichtspraxis*
ISBN 978-3-8325-1948-3 42.00 EUR

83 Corinna Kieren: Chemiehausaufgaben in der Sekundarstufe I des Gymnasiums. *Fragebogenerhebung zur gegenwärtigen Praxis und Entwicklung eines optimierten Hausaufgabendesigns im Themenbereich Säure-Base*
978-3-8325-1975-9 37.00 EUR

84 Marco Thiele: Modelle der Thermohalinen Zirkulation im Unterricht. *Eine empirische Studie zur Förderung des Modellverständnisses*
ISBN 978-3-8325-1982-7 40.50 EUR

85 Bernd Zinn: Physik lernen, um Physik zu lehren. *Eine Möglichkeit für interessanteren Physikunterricht*
ISBN 978-3-8325-1995-7 39.50 EUR

86 Esther Klaes: Außerschulische Lernorte im naturwissenschaftlichen Unterricht. *Die Perspektive der Lehrkraft*
ISBN 978-3-8325-2006-9 43.00 EUR

87 Marita Schmidt: Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. *Entwicklung und Erprobung eines Testinventars*
ISBN 978-3-8325-2024-3 37.00 EUR

88 Gudrun Franke-Braun: Aufgaben mit gestuften Lernhilfen. *Ein Aufgabenformat zur Förderung der sachbezogenen Kommunikation und Lernleistung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-2026-7 38.00 EUR

89 Silke Klos: Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht. *Der Einfluss eines integrierten Unterrichtskonzepts*
ISBN 978-3-8325-2133-2 37.00 EUR

90 Ulrike Elisabeth Burkard: Quantenphysik in der Schule. *Bestandsaufnahme, Perspektiven und Weiterentwicklungsmöglichkeiten durch die Implementation eines Medienservers*
ISBN 978-3-8325-2215-5 43.00 EUR

91 Ulrike Gromadecki: Argumente in physikalischen Kontexten. *Welche Geltungsgründe halten Physikanfänger für überzeugend?*
ISBN 978-3-8325-2250-6 41.50 EUR

92 Jürgen Bruns: Auf dem Weg zur Förderung naturwissenschaftsspezifischer Vorstellungen von zukünftigen Chemie-Lehrenden
ISBN 978-3-8325-2257-5 43.50 EUR

93 Cornelius Marsch: Räumliche Atomvorstellung. *Entwicklung und Erprobung eines Unterrichtskonzeptes mit Hilfe des Computers*
ISBN 978-3-8325-2293-3 82.50 EUR

94 Maja Brückmann: Sachstrukturen im Physikunterricht. *Ergebnisse einer Videostudie*
ISBN 978-3-8325-2272-8 39.50 EUR

95 Sabine Fechner: Effects of Context-oriented Learning on Student Interest and Achievement in Chemistry Education
ISBN 978-3-8325-2343-5 36.50 EUR

96 Clemens Nagel: eLearning im Physikalischen Anfängerpraktikum
ISBN 978-3-8325-2355-8 39.50 EUR

97 Josef Riese: Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften
ISBN 978-3-8325-2376-3 39.00 EUR

98 Sascha Bernholt: Kompetenzmodellierung in der Chemie. *Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität*
ISBN 978-3-8325-2447-0 40.00 EUR

99 Holger Christoph Stawitz: Auswirkung unterschiedlicher Aufgabenprofile auf die Schülerleistung. *Vergleich von Naturwissenschafts- und Problemlöseaufgaben der PISA 2003-Studie*
ISBN 978-3-8325-2451-7 37.50 EUR

100 Hans Ernst Fischer, Elke Sumfleth (Hrsg.): nwu-essen – 10 Jahre Essener Forschung zum naturwissenschaftlichen Unterricht
ISBN 978-3-8325-3331-1 40.00 EUR

101 Hendrik Härtig: Sachstrukturen von Physikschulbüchern als Grundlage zur Bestimmung der Inhaltsvalidität eines Tests
ISBN 978-3-8325-2512-5 34.00 EUR

102 Thomas Grüß-Niehaus: Zum Verständnis des Löslichkeitskonzeptes im Chemieunterricht. *Der Effekt von Methoden progressiver und kollaborativer Reflexion*
ISBN 978-3-8325-2537-8 40.50 EUR

103 Patrick Bronner: Quantenoptische Experimente als Grundlage eines Curriculums zur Quantenphysik des Photons
ISBN 978-3-8325-2540-8 36.00 EUR

104 Adrian Voßkühler: Blickbewegungsmessung an Versuchsaufbauten. *Studien zur Wahrnehmung, Verarbeitung und Usability von physikbezogenen Experimenten am Bildschirm und in der Realität*
ISBN 978-3-8325-2548-4 47.50 EUR

105 Verena Tobias: Newton'sche Mechanik im Anfangsunterricht. *Die Wirksamkeit einer Einführung über die zweidimensionale Dynamik auf das Lehren und Lernen*
ISBN 978-3-8325-2558-3 54.00 EUR

106 Christian Rogge: Entwicklung physikalischer Konzepte in aufgabenbasierten Lernumgebungen
ISBN 978-3-8325-2574-3 45.00 EUR

107 Mathias Ropohl: Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. *Entwicklung und Analyse von Testaufgaben*
ISBN 978-3-8325-2609-2 36.50 EUR

108 Christoph Kulgemeyer: Physikalische Kommunikationskompetenz. *Modellierung und Diagnostik*
ISBN 978-3-8325-2674-0 44.50 EUR

109 Jennifer Olszewski: The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes
ISBN 978-3-8325-2680-1 33.50 EUR

110 Annika Ohle: Primary School Teachers' Content Knowledge in Physics and its Impact on Teaching and Students' Achievement
ISBN 978-3-8325-2684-9 36.50 EUR

111 Susanne Mannel: Assessing scientific inquiry. *Development and evaluation of a test for the low-performing stage*
ISBN 978-3-8325-2761-7 40.00 EUR

112 Michael Plomer: Physik physiologisch passend praktiziert. *Eine Studie zur Lernwirksamkeit von traditionellen und adressatenspezifischen Physikpraktika für die Physiologie*
ISBN 978-3-8325-2804-1 34.50 EUR

113 Alexandra Schulz: Experimentierspezifische Qualitätsmerkmale im Chemieunterricht. *Eine Videostudie*
ISBN 978-3-8325-2817-1 40.00 EUR

114 Franz Boczianowski: Eine empirische Untersuchung zu Vektoren im Physikunterricht der Mittelstufe
ISBN 978-3-8325-2843-0 39.50 EUR

115 Maria Ploog: Internetbasiertes Lernen durch Textproduktion im Fach Physik
ISBN 978-3-8325-2853-9 39.50 EUR

116 Anja Dhein: Lernen in Explorier- und Experimentiersituationen. *Eine explorative Studie zu Bedeutungsentwicklungsprozessen bei Kindern im Alter zwischen 4 und 6 Jahren*
ISBN 978-3-8325-2859-1 45.50 EUR

117 Irene Neumann: Beyond Physics Content Knowledge. *Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge*
ISBN 978-3-8325-2880-5 37.00 EUR

118 Markus Emden: Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. *Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*
ISBN 978-3-8325-2867-6 38.00 EUR

119 Birgit Hofmann: Analyse von Blickbewegungen von Schülern beim Lesen von physikbezogenen Texten mit Bildern. *Eye Tracking als Methodenwerkzeug in der physikdidaktischen Forschung*
ISBN 978-3-8325-2925-3 59.00 EUR

120 Rebecca Knobloch: Analyse der fachinhaltlichen Qualität von Schüleräußerungen und deren Einfluss auf den Lernerfolg. *Eine Videostudie zu kooperativer Kleingruppenarbeit*
ISBN 978-3-8325-3006-8 36.50 EUR

121 Julia Hostenbach: Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht
ISBN 978-3-8325-3013-6 38.00 EUR

122 Anna Windt: Naturwissenschaftliches Experimentieren im Elementarbereich. *Evaluation verschiedener Lernsituationen*
ISBN 978-3-8325-3020-4 43.50 EUR

123 Eva Kölbach: Kontexteinflüsse beim Lernen mit Lösungsbeispielen
ISBN 978-3-8325-3025-9 38.50 EUR

124 Anna Lau: Passung und vertikale Vernetzung im Chemie- und Physikunterricht
ISBN 978-3-8325-3021-1 36.00 EUR

125 Jan Lamprecht: Ausbildungswege und Komponenten professioneller Handlungskompetenz. *Vergleich von Quereinsteigern mit Lehramtsabsolventen für Gymnasien im Fach Physik*
ISBN 978-3-8325-3035-8 38.50 EUR

126 Ulrike Böhm: Förderung von Verstehensprozessen unter Einsatz von Modellen
ISBN 978-3-8325-3042-6 41.00 EUR

127 Sabrina Dollny: Entwicklung und Evaluation eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Chemielehrkräften
ISBN 978-3-8325-3046-4 37.00 EUR

128 Monika Zimmermann: Naturwissenschaftliche Bildung im Kindergarten. *Eine integrative Längsschnittstudie zur Kompetenzentwicklung von Erzieherinnen*
ISBN 978-3-8325-3053-2 54.00 EUR

129 Ulf Saballus: Über das Schlussfolgern von Schülerinnen und Schülern zu öffentlichen Kontroversen mit naturwissenschaftlichem Hintergrund. *Eine Fallstudie*
ISBN 978-3-8325-3086-0 39.50 EUR

130 Olaf Krey: Zur Rolle der Mathematik in der Physik. *Wissenschaftstheoretische Aspekte und Vorstellungen Physiklernender*
ISBN 978-3-8325-3101-0 46.00 EUR

131 Angelika Wolf: Zusammenhänge zwischen der Eigenständigkeit im Physikunterricht, der Motivation, den Grundbedürfnissen und dem Lernerfolg von Schülern
ISBN 978-3-8325-3161-4 45.00 EUR

132 Johannes Börlin: Das Experiment als Lerngelegenheit. *Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität*
ISBN 978-3-8325-3170-6 45.00 EUR

133 Olaf Uhden: Mathematisches Denken im Physikunterricht. *Theorieentwicklung und Problemanalyse*
ISBN 978-3-8325-3170-6 45.00 EUR

134 Christoph Gut: Modellierung und Messung experimenteller Kompetenz. *Analyse eines large-scale Experimentiertests*
ISBN 978-3-8325-3213-0 40.00 EUR

135 Antonio Rueda: Lernen mit ExploMultimedial in kolumbianischen Schulen. *Analyse von kurzzeitigen Lernprozessen und der Motivation beim länderübergreifenden Einsatz einer deutschen computergestützten multimedialen Lernumgebung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-3218-5 45.50 EUR

136 Krisztina Berger: Bilder, Animationen und Notizen. *Empirische Untersuchung zur Wirkung einfacher visueller Repräsentationen und Notizen auf den Wissenserwerb in der Optik*
ISBN 978-3-8325-3238-3 41.50 EUR

137 Antony Crossley: Untersuchung des Einflusses unterschiedlicher physikalischer Konzepte auf den Wissenserwerb in der Thermodynamik der Sekundarstufe I
ISBN 978-3-8325-3275-8 40.00 EUR

138 Tobias Viering: Entwicklung physikalischer Kompetenz in der Sekundarstufe I. *Validierung eines Kompetenzentwicklungsmodells für das Energiekonzept im Bereich Fachwissen*
ISBN 978-3-8325-3277-2 37.00 EUR

139 Nico Schreiber: Diagnostik experimenteller Kompetenz. *Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*
ISBN 978-3-8325-3284-0 39.00 EUR

140 Sarah Hundertmark: Einblicke in kollaborative Lernprozesse. *Eine Fallstudie zur reflektierenden Zusammenarbeit unterstützt durch die Methoden Concept Mapping und Lernbegleitbogen*
ISBN 978-3-8325-3251-2 43.00 EUR

141 Ronny Scherer: Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie. *Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II*
ISBN 978-3-8325-3312-0 43.00 EUR

142 Patricia Heitmann: Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. *Modellierung und Diagnose der Kompetenzen Bewertung und analytisches Problemlösen für das Fach Chemie*
ISBN 978-3-8325-3314-4 37.00 EUR

143 Jan Fleischhauer: Wissenschaftliches Argumentieren und Entwicklung von Konzepten beim Lernen von Physik
ISBN 978-3-8325-3325-0 35.00 EUR

144 Nermin Özcan: Zum Einfluss der Fachsprache auf die Leistung im Fach Chemie. *Eine Förderstudie zur Fachsprache im Chemieunterricht*
ISBN 978-3-8325-3328-1 36.50 EUR

145 Helena van Vorst: Kontextmerkmale und ihr Einfluss auf das Schülerinteresse im Fach Chemie
ISBN 978-3-8325-3321-2 38.50 EUR

146 Janine Cappell: Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase
ISBN 978-3-8325-3356-4 38.50 EUR

147 Susanne Bley: Förderung von Transferprozessen im Chemieunterricht
ISBN 978-3-8325-3407-3 40.50 EUR

148 Cathrin Blaes: Die übungsgestützte Lehrerpräsentation im Chemieunterricht der Sekundarstufe I. *Evaluation der Effektivität*
ISBN 978-3-8325-3409-7 43.50 EUR

149 Julia Suckut: Die Wirksamkeit von piko-OWL als Lehrerfortbildung. Eine Evaluation zum Projekt *Physik im Kontext* in Fallstudien
ISBN 978-3-8325-3440-0 45.00 EUR

150 Alexandra Dorschu: Die Wirkung von Kontexten in Physikkompetenztestaufgaben
ISBN 978-3-8325-3446-2 37.00 EUR

151 Jochen Scheid: Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung: *Ein Beitrag zur Entwicklung der Aufgabenkultur*
ISBN 978-3-8325-3449-3 49.00 EUR

152 Tim Plasa: Die Wahrnehmung von Schülerlaboren und Schülerforschungszentren
ISBN 978-3-8325-3483-7 35.50 EUR

153 Felix Schoppmeier: Physikkompetenz in der gymnasialen Oberstufe.*Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*
ISBN 978-3-8325-3502-5 36.00 EUR

154 Katharina Groß: Experimente alternativ dokumentieren. *Eine qualitative Studie zur Förderung der Diagnose- und Differenzierungskompetenz in der Chemielehrerbildung*
ISBN 978-3-8325-3508-7 43.50 EUR

155 Barbara Hank: Konzeptwandelprozesse im Anfangsunterricht Chemie. *Eine quasiexperimentelle Längsschnittstudie*
ISBN 978-3-8325-3519-3 38.50 EUR

156 Katja Freyer: Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie
ISBN 978-3-8325-3544-5 38.00 EUR

157 Alexander Rachel: Auswirkungen instruktionaler Hilfen bei der Einführung des (Ferro-)Magnetismus. *Eine Vergleichsstudie in der Primar- und Sekundarstufe*
ISBN 978-3-8325-3548-3 43.50 EUR

158 Sebastian Ritter: Einfluss des Lerninhalts Nanogrößeneffekte auf Teilchen- und Teilchenmodellvorstellungen von Schülerinnen und Schülern
ISBN 978-3-8325-3558-2 36.00 EUR

159 Andrea Harbach: Problemorientierung und Vernetzung in kontextbasierten Lernaufgaben
ISBN 978-3-8325-3564-3 39.00 EUR

160 David Obst: Interaktive Tafeln im Physikunterricht. *Entwicklung und Evaluation einer Lehrerfortbildung*
ISBN 978-3-8325-3582-7 40.50 EUR

161 Sophie Kirschner: Modellierung und Analyse des Professionswissens von Physiklehrkräften
ISBN 978-3-8325-3601-5 35.00 EUR

162 Katja Stief: Selbstregulationsprozesse und Hausaufgabenmotivation im Chemieunterricht
ISBN 978-3-8325-3631-2 34.00 EUR

163 Nicola Meschede: Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. *Theoretische Beschreibung und empirische Erfassung*
ISBN 978-3-8325-3668-8 37.00 EUR

164 Johannes Maximilian Barth: Experimentieren im Physikunterricht der gymnasialen Oberstufe. *Eine Rekonstruktion übergeordneter Einbettungsstrategien*
ISBN 978-3-8325-3681-7 39.00 EUR

165 Sandra Lein: Das Betriebspraktikum in der Lehrerbildung. *Eine Untersuchung zur Förderung der Wissenschafts- und Technikbildung im allgemeinbildenden Unterricht*
ISBN 978-3-8325-3698-5 40.00 EUR

166 Veranika Maiseyenka: Modellbasiertes Experimentieren im Unterricht. *Praxistauglichkeit und Lernwirkungen*
ISBN 978-3-8325-3708-1 38.00 EUR

167 Christoph Stolzenberger: Der Einfluss der didaktischen Lernumgebung auf das Erreichen geforderter Bildungsziele am Beispiel der W- und P-Seminare im Fach Physik
ISBN 978-3-8325-3708-1 38.00 EUR

168 Pia Altenburger: Mehrebenenregressionsanalysen zum Physiklernen im Sachunterricht der Primarstufe. *Ergebnisse einer Evaluationsstudie.*
ISBN 978-3-8325-3717-3 37.50 EUR

169 Nora Ferber: Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I
ISBN 978-3-8325-3727-2 39.50 EUR

170 Anita Stender: Unterrichtsplanung: Vom Wissen zum Handeln.
Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung
ISBN 978-3-8325-3750-0 41.50 EUR

171 Jenna Koenen: Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen
ISBN 978-3-8325-3785-2 43.00 EUR

172 Teresa Henning: Empirische Untersuchung kontextorientierter Lernumgebungen in der Hochschuldidaktik. *Entwicklung und Evaluation kontextorientierter Aufgaben in der Studieneingangsphase für Fach- und Nebenfachstudierende der Physik*
ISBN 978-3-8325-3801-9 43.00 EUR

173 Alexander Pusch: Fachspezifische Instrumente zur Diagnose und individuellen Förderung von Lehramtsstudierenden der Physik
ISBN 978-3-8325-3829-3 38.00 EUR

174 Christoph Vogelsang: Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. *Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*
ISBN 978-3-8325-3846-0 50.50 EUR

175 Ingo Brebeck: Selbstreguliertes Lernen in der Studieneingangsphase im Fach Chemie
ISBN 978-3-8325-3859-0 37.00 EUR

176 Axel Eghtessad: Merkmale und Strukturen von Professionalisierungsprozessen in der ersten und zweiten Phase der Chemielehrerbildung. *Eine empirisch-qualitative Studie mit niedersächsischen Fachleiter_innen der Sekundarstufenlehrämter*
ISBN 978-3-8325-3861-3 45.00 EUR

177 Andreas Nehring: Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung
ISBN 978-3-8325-3872-9 39.50 EUR

178 Maike Schmidt: Professionswissen von Sachunterrichtslehrkräften. Zusammenhangsanalyse zur Wirkung von Ausbildungshintergrund und Unterrichtserfahrung auf das fachspezifische Professionswissen im Unterrichtsinhalt „Verbrennung“
ISBN 978-3-8325-3907-8 38.50 EUR

179 Jan Winkelmann: Auswirkungen auf den Fachwissenszuwachs und auf affektive Schülermerkmale durch Schüler- und Demonstrationsexperimente im Physikunterricht
ISBN 978-3-8325-3915-3 41.00 EUR

180 Iwen Kobow: Entwicklung und Validierung eines Testinstrumentes zur Erfassung der Kommunikationskompetenz im Fach Chemie
ISBN 978-3-8325-3927-6 34.50 EUR

181 Yvonne Gramzow: Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion
ISBN 978-3-8325-3931-3 42.50 EUR

182 Evelin Schröter: Entwicklung der Kompetenzerwartung durch Lösen physikalischer Aufgaben einer multimedialen Lernumgebung
ISBN 978-3-8325-3975-7 54.50 EUR

183 Inga Kallweit: Effektivität des Einsatzes von Selbsteinschätzungsbögen im Chemieunterricht der Sekundarstufe I. *Individuelle Förderung durch selbstreguliertes Lernen*
ISBN 978-3-8325-3965-8 44.00 EUR

184 Andrea Schumacher: Paving the way towards authentic chemistry teaching. *A contribution to teachers' professional development*
ISBN 978-3-8325-3976-4 48.50 EUR

185 David Woitkowski: Fachliches Wissen Physik in der Hochschulausbildung. *Konzeptualisierung, Messung, Niveaubildung*
ISBN 978-3-8325-3988-7 53.00 EUR

186 Marianne Korner: Cross-Age Peer Tutoring in Physik. *Evaluation einer Unterrichtsmethode*
ISBN 978-3-8325-3979-5 38.50 EUR

187 Simone Nakoinz: Untersuchung zur Verknüpfung submikroskopischer und makroskopischer Konzepte im Fach Chemie
ISBN 978-3-8325-4057-9 38.50 EUR

188 Sandra Anus: Evaluation individueller Förderung im Chemieunterricht.*Adaptivität von Lerninhalten an das Vorwissen von Lernenden am Beispiel des Basiskonzeptes Chemische Reaktion*
ISBN 978-3-8325-4059-3 43.50 EUR

189 Thomas Roßbegalle: Fachdidaktische Entwicklungsforschung zum besseren Verständnis atmosphärischer Phänomene. *Treibhauseffekt, saurer Regen und stratosphärischer Ozonabbau als Kontexte zur Vermittlung von Basiskonzepten der Chemie*
ISBN 978-3-8325-4059-3 45.50 EUR

190 Kathrin Steckenmesser-Sander: Gemeinsamkeiten und Unterschiede physikbezogener Handlungs-, Denk- und Lernprozesse von Mädchen und Jungen
ISBN 978-3-8325-4066-1 38.50 EUR

191 Cornelia Geller: Lernprozessorientierte Sequenzierung des Physikunterrichts im Zusammenhang mit Fachwissenserwerb. *Eine Videostudie in Finnland, Deutschland und der Schweiz*
ISBN 978-3-8325-4082-1 35.50 EUR

192 Jan Hofmann: Untersuchung des Kompetenzaufbaus von Physiklehrkräften während einer Fortbildungsmaßnahme
ISBN 978-3-8325-4104-0 38.50 EUR

193 Andreas Dickhäuser: Chemiespezifischer Humor. *Theoriebildung, Materialentwicklung, Evaluation*
ISBN 978-3-8325-4108-8 37.00 EUR

194 Stefan Korte: Die Grenzen der Naturwissenschaft als Thema des Physikunterrichts
ISBN 978-3-8325-4112-5 57.50 EUR

195 Carolin Hülsmann: Kurswahlmotive im Fach Chemie. Eine Studie zum Wahlverhalten und Erfolg von Schülerinnen und Schülern in der gymnasialen Oberstufe
ISBN 978-3-8325-4144-6 49.00 EUR

196 Caroline Körbs: Mindeststandards im Fach Chemie am Ende der Pflichtschulzeit
ISBN 978-3-8325-4148-4 34.00 EUR

197 Andreas Vorholzer: Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? *Eine empirische Untersuchung der Wirkung eines expliziten und eines impliziten Instruktionsansatzes*
ISBN 978-3-8325-4194-1 37.50 EUR

198 Anna Katharina Schmitt: Entwicklung und Evaluation einer Chemielehrerfortbildung zum Kompetenzbereich Erkenntnisgewinnung
ISBN 978-3-8325-4228-3 39.50 EUR

199 Christian Maurer: Strukturierung von Lehr-Lern-Sequenzen
ISBN 978-3-8325-4247-4 36.50 EUR

200 Helmut Fischler, Elke Sumfleth (Hrsg.): Professionelle Kompetenz von Lehrkräften der Chemie und Physik
ISBN 978-3-8325-4523-9 34.00 EUR

201 Simon Zander: Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen
ISBN 978-3-8325-4248-1 35.00 EUR

202 Kerstin Arndt: Experimentierkompetenz erfassen.
Analyse von Prozessen und Mustern am Beispiel von Lehramtsstudierenden der Chemie
ISBN 978-3-8325-4266-5 45.00 EUR

203 Christian Lang: Kompetenzorientierung im Rahmen experimentalchemischer Praktika
ISBN 978-3-8325-4268-9 42.50 EUR

204 Eva Cauet: Testen wir relevantes Wissen? *Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*
ISBN 978-3-8325-4276-4 39.50 EUR

205 Patrick Löffler: Modellanwendung in Problemlöseaufgaben. *Wie wirkt Kontext?*
ISBN 978-3-8325-4303-7 35.00 EUR

206 Carina Gehlen: Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie
ISBN 978-3-8325-4318-1 43.00 EUR

207 Lars Oettinghaus: Lehrerüberzeugungen und physikbezogenes Professionswissen. *Vergleich von Absolventinnen und Absolventen verschiedener Ausbildungswege im Physikreferendariat*
ISBN 978-3-8325-4319-8 38.50 EUR

208 Jennifer Petersen: Zum Einfluss des Merkmals Humor auf die Gesundheitsförderung im Chemieunterricht der Sekundarstufe I.
Eine Interventionsstudie zum Thema Sonnenschutz
ISBN 978-3-8325-4348-8 40.00 EUR

209 Philipp Straube: Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik
ISBN 978-3-8325-4351-8 35.50 EUR

210 Martin Dickmann: Messung von Experimentierfähigkeiten. *Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*
ISBN 978-3-8325-4356-3 41.00 EUR

211 Markus Bohlmann: Science Education. Empirie, Kulturen und Mechanismen der Didaktik der Naturwissenschaften
ISBN 978-3-8325-4377-8 44.00 EUR

212 Martin Draude: Die Kompetenz von Physiklehrkräften, Schwierigkeiten von Schülerinnen und Schülern beim eigenständigen Experimentieren zu diagnostizieren
ISBN 978-3-8325-4382-2 37.50 EUR

213 Henning Rode: Prototypen evidenzbasierten Physikunterrichts. *Zwei empirische Studien zum Einsatz von Feedback und Blackboxes in der Sekundarstufe*
ISBN 978-3-8325-4389-1 42.00 EUR

214 Jan-Henrik Kechel: Schülerschwierigkeiten beim eigenständigen Experimentieren. *Eine qualitative Studie am Beispiel einer Experimentieraufgabe zum Hooke'schen Gesetz*
ISBN 978-3-8325-4392-1 55.00 EUR

215 Katharina Fricke: Classroom Management and its Impact on Lesson Outcomes in Physics. *A multi-perspective comparison of teaching practices in primary and secondary schools*
ISBN 978-3-8325-4394-5 40.00 EUR

216 Hannes Sander: Orientierungen von Jugendlichen beim Urteilen und Entscheiden in Kontexten nachhaltiger Entwicklung. *Eine rekonstruktive Perspektive auf Bewertungskompetenz in der Didaktik der Naturwissenschaft*
ISBN 978-3-8325-4434-8 46.00 EUR

217 Inka Haak: Maßnahmen zur Unterstützung kognitiver und metakognitiver Prozesse in der Studieneingangsphase. *Eine Design-Based-Research-Studie zum universitären Lernzentrum Physiktreff*
ISBN 978-3-8325-4437-9 46.50 EUR

218 Martina Brandenburger: Was beeinflusst den Erfolg beim Problemlösen in der Physik? *Eine Untersuchung mit Studierenden*
ISBN 978-3-8325-4409-6 42.50 EUR

219 Corinna Helms: Entwicklung und Evaluation eines Trainings zur Verbesserung der Erklärqualität von Schülerinnen und Schülern im Gruppenpuzzle
ISBN 978-3-8325-4454-6 42.50 EUR

220 Viktoria Rath: Diagnostische Kompetenz von angehenden Physiklehrkräften. *Modellierung, Testinstrumentenentwicklung und Erhebung der Performanz bei der Diagnose von Schülervorstellungen in der Mechanik*
ISBN 978-3-8325-4456-0 42.50 EUR

221 Janne Krüger: Schülerperspektiven auf die zeitliche Entwicklung der Naturwissenschaften
ISBN 978-3-8325-4457-7 45.50 EUR

222 Stefan Mutke: Das Professionswissen von Chemiereferendarinnen und -referendaren in Nordrhein-Westfalen. *Eine Längsschnittstudie*
ISBN 978-3-8325-4458-4 37.50 EUR

223 Sebastian Habig: Systematisch variierte Kontextaufgaben und ihr Einfluss auf kognitive und affektive Schülerfaktoren
ISBN 978-3-8325-4467-6 40.50 EUR

224 Sven Liepertz: Zusammenhang zwischen dem Professionswissen von Physiklehrkräften, dem sachstrukturellen Angebot des Unterrichts und der Schülerleistung
ISBN 978-3-8325-4480-5 34.00 EUR

225 Elina Platova: Optimierung eines Laborpraktikums durch kognitive Aktivierung
ISBN 978-3-8325-4481-2 39.00 EUR

226 Tim Reschke: Lesegeschichten im Chemieunterricht der Sekundarstufe I zur Unterstützung von situationalem Interesse und Lernerfolg
ISBN 978-3-8325-4487-4 41.00 EUR

227 Lena Mareike Walper: Entwicklung der physikbezogenen Interessen und selbstbezogenen Kognitionen von Schülerinnen und Schülern in der Übergangsphase von der Primar- in die Sekundarstufe. *Eine Längsschnittanalyse vom vierten bis zum siebten Schuljahr*
ISBN 978-3-8325-4495-9 43.00 EUR

228 Stefan Anthofer: Förderung des fachspezifischen Professionswissens von Chemielehramtsstudierenden
ISBN 978-3-8325-4498-0 39.50 EUR

229 Marcel Bullinger: Handlungsorientiertes Physiklernen mit instruierten Selbsterklärungen in der Primarstufe. *Eine experimentelle Laborstudie*
ISBN 978-3-8325-4504-8 44.00 EUR

230 Thomas Amenda: Bedeutung fachlicher Elementarisierungen für das Verständnis der Kinematik
ISBN 978-3-8325-4531-4 43.50 EUR

231 Sabrina Milke: Beeinflusst *Priming* das Physiklernen?
Eine empirische Studie zum Dritten Newtonschen Axiom
ISBN 978-3-8325-4549-4 42.00 EUR

232 Corinna Erfmann: Ein anschaulicher Weg zum Verständnis der elektromagnetischen Induktion. *Evaluation eines Unterrichtsvorschlags und Validierung eines Leistungsdiagnoseinstruments*
ISBN 978-3-8325-4550-5 49.50 EUR

233 Hanne Rautenstrauch: Erhebung des (Fach-)Sprachstandes bei Lehramtsstudierenden im Kontext des Faches Chemie
ISBN 978-3-8325-4556-7 40.50 EUR

234 Tobias Klug: Wirkung kontextorientierter physikalischer Praktikumsversuche auf Lernprozesse von Studierenden der Medizin
ISBN 978-3-8325-4558-1 37.00 EUR

235 Mareike Bohrmann: Zur Förderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht
ISBN 978-3-8325-4559-8 52.00 EUR

236 Anja Schödl: FALKO-Physik – Fachspezifische Lehrerkompetenzen im Fach Physik. *Entwicklung und Validierung eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Physiklehrkräften*
ISBN 978-3-8325-4553-6 40.50 EUR

237 Hilda Scheuermann: Entwicklung und Evaluation von Unterstützungsmaßnahmen zur Förderung der Variablenkontrollstrategie beim Planen von Experimenten
ISBN 978-3-8325-4568-0 39.00 EUR

238 Christian G. Strippel: Naturwissenschaftliche Erkenntnisgewinnung an chemischen Inhalten vermitteln. *Konzeption und empirische Untersuchung einer Ausstellung mit Experimentierstation*
ISBN 978-3-8325-4577-2 41.50 EUR

239 Sarah Rau: Durchführung von Sachunterricht im Vorbereitungsdienst. *Eine längsschnittliche, videobasierte Unterrichtsanalyse*
ISBN 978-3-8325-4579-6 46.00 EUR

240 Thomas Plotz: Lernprozesse zu nicht-sichtbarer Strahlung. *Empirische Untersuchungen in der Sekundarstufe 2*
ISBN 978-3-8325-4624-3 39.50 EUR

241 Wolfgang Aschauer: Elektrische und magnetische Felder. *Eine empirische Studie zu Lernprozessen in der Sekundarstufe II*
ISBN 978-3-8325-4625-0 50.00 EUR

242 Anna Donhauser: Didaktisch rekonstruierte Materialwissenschaft. *Aufbau und Konzeption eines Schülerlabors für den Exzellenzcluster Engineering of Advanced Materials*
ISBN 978-3-8325-4636-6 39.00 EUR

243 Katrin Schüßler: Lernen mit Lösungsbeispielen im Chemieunterricht. *Einflüsse auf Lernerfolg, kognitive Belastung und Motivation*
ISBN 978-3-8325-4640-3 42.50 EUR

244 Timo Fleischer: Untersuchung der chemischen Fachsprache unter besonderer Berücksichtigung chemischer Repräsentationen
ISBN 978-3-8325-4642-7 46.50 EUR

245 Rosina Steininger: Concept Cartoons als Stimuli für Kleingruppendiskussionen im Chemieunterricht. *Beschreibung und Analyse einer komplexen Lerngelegenheit*
ISBN 978-3-8325-4647-2 39.00 EUR

246 Daniel Rehfeldt: Erfassung der Lehrqualität naturwissenschaftlicher Experimentalpraktika
ISBN 978-3-8325-4590-1 40.00 EUR

247 Sandra Puddu: Implementing Inquiry-based Learning in a Diverse Classroom: Investigating Strategies of Scaffolding and Students' Views of Scientific Inquiry
ISBN 978-3-8325-4591-8 35.50 EUR

248 Markus Bliersbach: Kreativität in der Chemie. *Erhebung und Förderung der Vorstellungen von Chemielehramtsstudierenden*
ISBN 978-3-8325-4593-2 44.00 EUR

249 Lennart Kimpel: Aufgaben in der Allgemeinen Chemie. *Zum Zusammenspiel von chemischem Verständnis und Rechenfähigkeit*
ISBN 978-3-8325-4618-2 36.00 EUR

250 Louise Bindel: Effects of integrated learning: explicating a mathematical concept in inquiry-based science camps
ISBN 978-3-8325-4655-7 37.50 EUR

251 Michael Wenzel: Computereinsatz in Schule und Schülerlabor. *Einstellung von Physiklehrkräften zu Neuen Medien*
ISBN 978-3-8325-4659-5 38.50 EUR

252 Laura Muth: Einfluss der Auswertephase von Experimenten im Physikunterricht. *Ergebnisse einer Interventionsstudie zum Zuwachs von Fachwissen und experimenteller Kompetenz von Schülerinnen und Schülern*
ISBN 978-3-8325-4675-5 36.50 EUR

253 Annika Fricke: Interaktive Skripte im Physikalischen Praktikum. *Entwicklung und Evaluation von Hypermedien für die Nebenfachausbildung*
ISBN 978-3-8325-4676-2 41.00 EUR

254 Julia Haase: Selbstbestimmtes Lernen im naturwissenschaftlichen Sachunterricht. *Eine empirische Interventionsstudie mit Fokus auf Feedback und Kompetenzerleben*
ISBN 978-3-8325-4685-4 38.50 EUR

255 Antje J. Heine: Was ist Theoretische Physik? *Eine wissenschaftstheoretische Betrachtung und Rekonstruktion von Vorstellungen von Studierenden und Dozenten über das Wesen der Theoretischen Physik*
ISBN 978-3-8325-4691-5 46.50 EUR

256 Claudia Meinhardt: Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern
ISBN 978-3-8325-4712-7 47.00 EUR

257 Ann-Kathrin Schlüter: Professionalisierung angehender Chemielehrkräfte für einen Gemeinsamen Unterricht
ISBN 978-3-8325-4713-4 53.50 EUR

258 Stefan Richtberg: Elektronenbahnen in Feldern. Konzeption und Evaluation einer webbasierten Lernumgebung
ISBN 978-3-8325-4723-3 49.00 EUR

259 Jan-Philipp Burde: Konzeption und Evaluation eines Unterrichtskonzepts zu einfachen Stromkreisen auf Basis des Elektronengasmodells
ISBN 978-3-8325-4726-4 57.50 EUR

260 Frank Finkenberg: Flipped Classroom im Physikunterricht
ISBN 978-3-8325-4737-4 42.50 EUR

261 Florian Treisch: Die Entwicklung der Professionellen Unterrichtswahrnehmung im Lehr-Lern-Labor Seminar
ISBN 978-3-8325-4741-4 41.50 EUR

262 Desiree Mayr: Strukturiertheit des experimentellen naturwissenschaftlichen Problemlöseprozesses
ISBN 978-3-8325-4757-8 37.00 EUR

263 Katrin Weber: Entwicklung und Validierung einer Learning Progression für das Konzept der chemischen Reaktion in der Sekundarstufe I
ISBN 978-3-8325-4762-2 48.50 EUR

264 Hauke Bartels: Entwicklung und Bewertung eines performanznahen Videovignettentests zur Messung der Erklärfähigkeit von Physiklehrkräften
ISBN 978-3-8325-4804-9 37.00 EUR

265 Karl Marniok: Zum Wesen von Theorien und Gesetzen in der Chemie. *Begriffsanalyse und Förderung der Vorstellungen von Lehramtsstudierenden*
ISBN 978-3-8325-4805-6 42.00 EUR

266 Marisa Holzapfel: Fachspezifischer Humor als Methode in der Gesundheitsbildung im Übergang von der Primarstufe zur Sekundarstufe I
ISBN 978-3-8325-4808-7 50.00 EUR

267 Anna Stolz: Die Auswirkungen von Experimentiersituationen mit unterschiedlichem Öffnungsgrad auf Leistung und Motivation der Schülerinnen und Schüler
ISBN 978-3-8325-4781-3 38.00 EUR

268 Nina Ulrich: Interaktive Lernaufgaben in dem digitalen Schulbuch eChemBook. *Einfluss des Interaktivitätsgrads der Lernaufgaben und des Vorwissens der Lernenden auf den Lernerfolg*
ISBN 978-3-8325-4814-8 43.50 EUR

269 Kim-Alessandro Weber: Quantenoptik in der Lehrerfortbildung. *Ein bedarfsgeprägtes Fortbildungskonzept zum Quantenobjekt „Photon“ mit Realexperimenten*
ISBN 978-3-8325-4792-9 55.00 EUR

270 Nina Skorsetz: Empathisierer und Systematisierer im Vorschulalter. *Eine Fragebogen- und Videostudie zur Motivation, sich mit Naturphänomenen zu beschäftigen*
ISBN 978-3-8325-4825-4 43.50 EUR

271 Franziska Kehne: Analyse des Transfers von kontextualisiert erworbenem Wissen im Fach Chemie
ISBN 978-3-8325-4846-9 45.00 EUR

272 Markus Elsholz: Das akademische Selbstkonzept angehender Physiklehrkräfte als Teil ihrer professionellen Identität. *Dimensionalität und Veränderung während einer zentralen Praxisphase*
ISBN 978-3-8325-4857-5 37.50 EUR

273 Joachim Müller: Studienerfolg in der Physik. *Zusammenhang zwischen Modellierungskompetenz und Studienerfolg*
ISBN 978-3-8325-4859-9 35.00 EUR

274 Jennifer Dörschelln: Organische Leuchtdioden. *Implementation eines innovativen Themas in den Chemieunterricht*
ISBN 978-3-8325-4865-0 59.00 EUR

275 Stephanie Strelow: Beliefs von Studienanfängern des Kombi-Bachelors Physik über die Natur der Naturwissenschaften
ISBN 978-3-8325-4881-0 40.50 EUR

276 Dennis Jaeger: Kognitive Belastung und aufgabenspezifische sowie personenspezifische Einflussfaktoren beim Lösen von Physikaufgaben
ISBN 978-3-8325-4928-2 50.50 EUR

277 Vanessa Fischer: Der Einfluss von Interesse und Motivation auf die Messung von Fach- und Bewertungskompetenz im Fach Chemie
ISBN 978-3-8325-4933-6 39.00 EUR

278 René Dohrmann: Professionsbezogene Wirkungen einer Lehr-Lern-Labor-Veranstaltung. *Eine multimethodische Studie zu den professionsbezogenen Wirkungen einer Lehr-Lern-Labor-Blockveranstaltung auf Studierende der Bachelorstudiengänge Lehramt Physik und Grundschulpädagogik (Sachunterricht)*
ISBN 978-3-8325-4958-9 40.00 EUR

279 Meike Bergs: Can We Make Them Use These Strategies? *Fostering Inquiry-Based Science Learning Skills with Physical and Virtual Experimentation Environments*
ISBN 978-3-8325-4962-6 39.50 EUR

280 Marie-Therese Hauerstein: Untersuchung zur Effektivität von Strukturierung und Binnendifferenzierung im Chemieunterricht der Sekundarstufe I. *Evaluation der Strukturierungshilfe Lernleiter*
ISBN 978-3-8325-4982-4 42.50 EUR

281 Verena Zucker: Erkennen und Beschreiben von formativem Assessment im naturwissenschaftlichen Grundschulunterricht. *Entwicklung eines Instruments zur Erfassung von Teilfähigkeiten der professionellen Wahrnehmung von Lehramtsstudierenden*
ISBN 978-3-8325-4991-6 38.00 EUR

282 Victoria Telser: Erfassung und Förderung experimenteller Kompetenz von Lehrkräften im Fach Chemie
ISBN 978-3-8325-4996-1 50.50 EUR

283 Kristine Tschirschky: Entwicklung und Evaluation eines gedächtnisorientierten Aufgabendesigns für Physikaufgaben
ISBN 978-3-8325-5002-8 42.50 EUR

284 Thomas Elert: Course Success in the Undergraduate General Chemistry Lab
ISBN 978-3-8325-5004-2 41.50 EUR

285 Britta Kalthoff: Explizit oder implizit? *Untersuchung der Lernwirksamkeit verschiedener fachmethodischer Instruktionen im Hinblick auf fachmethodische und fachinhaltliche Fähigkeiten von Sachunterrichtsstudierenden*
ISBN 978-3-8325-5013-4 37.50 EUR

286 Thomas Dickmann: Visuelles Modellverständnis und Studienerfolg in der Chemie. *Zwei Seiten einer Medaille*
ISBN 978-3-8325-5016-5 44.00 EUR

287 Markus Sebastian Feser: Physiklehrkräfte korrigieren Schülertexte. *Eine Explorationsstudie zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung im Physikunterricht*
ISBN 978-3-8325-5020-2 49.00 EUR

288 Matylda Dudzinska: Lernen mit Beispielaufgaben und Feedback im Physikunterricht der Sekundarstufe 1. *Energieerhaltung zur Lösung von Aufgaben nutzen*
ISBN 978-3-8325-5025-7 47.00 EUR

289 Ines Sonnenschein: Naturwissenschaftliche Denk- und Arbeitsprozesse Studierender im Labor
ISBN 978-3-8325-5033-2 52.00 EUR

290 Florian Simon: Der Einfluss von Betreuung und Betreuenden auf die Wirksamkeit von Schülerlaborbesuchen. *Eine Zusammenhangsanalyse von Betreuungsqualität, Betreuermerkmalen und Schülerlaborzielen sowie Replikationsstudie zur Wirksamkeit von Schülerlaborbesuchen*
ISBN 978-3-8325-5036-3 49.50 EUR

291 Marie-Annette Geyer: Physikalisch-mathematische Darstellungswechsel funktionaler Zusammenhänge. *Das Vorgehen von SchülerInnen der Sekundarstufe 1 und ihre Schwierigkeiten*
ISBN 978-3-8325-5047-9 46.50 EUR

292 Susanne Digel: Messung von Modellierungskompetenz in Physik. *Theoretische Herleitung und empirische Prüfung eines Kompetenzmodells physikspezifischer Modellierungskompetenz*
ISBN 978-3-8325-5055-4 41.00 EUR

293 Sönke Janssen: Angebots-Nutzungs-Prozesse eines Schülerlabors analysieren und gestalten. *Ein design-based research Projekt*
ISBN 978-3-8325-5065-3 57.50 EUR

294 Knut Wille: Der Productive Failure Ansatz als Beitrag zur Weiterentwicklung der Aufgabenkultur
ISBN 978-3-8325-5074-5 49.00 EUR

295 Lisanne Kraeva: Problemlösestrategien von Schülerinnen und Schülern diagnostizieren
ISBN 978-3-8325-5110-0 59.50 EUR

296 Jenny Lorentzen: Entwicklung und Evaluation eines Lernangebots im Lehramtsstudium Chemie zur Förderung von Vernetzungen innerhalb des fachbezogenen Professionswissens
ISBN 978-3-8325-5120-9 39.50 EUR

297 Micha Winkelmann: Lernprozesse in einem Schülerlabor unter Berücksichtigung individueller naturwissenschaftlicher Interessenstrukturen
ISBN 978-3-8325-5147-6 48.50 EUR

298 Carina Wöhlke: Entwicklung und Validierung eines Instruments zur Erfassung der professionellen Unterrichtswahrnehmung angehender Physiklehrkräfte
ISBN 978-3-8325-5149-0 43.00 EUR

299 Thomas Schubatzky: Das Amalgam Anfangs-Elektrizitätslehreunterricht. *Eine multiperspektivische Betrachtung in Deutschland und Österreich*
ISBN 978-3-8325-5159-9 50.50 EUR

300 Amany Annaggar: A Design Framework for Video Game-Based Gamification Elements to Assess Problem-solving Competence in Chemistry Education
ISBN 978-3-8325-5150-6 52.00 EUR

301 Alexander Engl: Chemie Pur – Unterrichten in der Natur: *Entwicklung und Evaluation eines kontextorientierten Unterrichtskonzepts im Bereich Outdoor Education zur Änderung der Einstellung zu „Chemie und Natur“*
ISBN 978-3-8325-5174-2 59.00 EUR

302 Christin Marie Sajons: Kognitive und motivationale Dynamik in Schülerlaboren. *Kontextualisierung, Problemorientierung und Autonomieunterstützung der didaktischen Struktur analysieren und weiterentwickeln*
ISBN 978-3-8325-5155-1 56.00 EUR

303 Philipp Bitzenbauer: Quantenoptik an Schulen. *Studie im Mixed-Methods Design zur Evaluation des Erlanger Unterrichtskonzepts zur Quantenoptik*
ISBN 978-3-8325-5123-0 59.00 EUR

304 Malte S. Ubben: Typisierung des Verständnisses mentaler Modelle mittels empirischer Datenerhebung am Beispiel der Quantenphysik
ISBN 978-3-8325-5181-0 43.50 EUR

305 Wiebke Kuske-Janßen: Sprachlicher Umgang mit Formeln von LehrerInnen im Physikunterricht am Beispiel des elektrischen Widerstandes in Klassenstufe 8
ISBN 978-3-8325-5183-4 47.50 EUR

306 Kai Bliesmer: Physik der Küste für außerschulische Lernorte. *Eine Didaktische Rekonstruktion*
ISBN 978-3-8325-5190-2 58.00 EUR

307 Nikola Schild: Eignung von domänenspezifischen Studieneingangsvariablen als Prädiktoren für Studienerfolg im Fach und Lehramt Physik
ISBN 978-3-8325-5226-8 42.00 EUR

308 Daniel Averbeck: Zum Studienerfolg in der Studieneingangsphase des Chemiestudiums. *Der Einfluss kognitiver und affektiv-motivationaler Variablen*
ISBN 978-3-8325-5227-5 51.00 EUR

309 Martina Strübe: Modelle und Experimente im Chemieunterricht. *Eine Videostudie zum fachspezifischen Lehrerwissen und -handeln*
ISBN 978-3-8325-5245-9 45.50 EUR

310 Wolfgang Becker: Auswirkungen unterschiedlicher experimenteller Repräsentationen auf den Kenntnisstand bei Grundschulkindern
ISBN 978-3-8325-5255-8 50.00 EUR

311 Marvin Rost: Modelle als Mittel der Erkenntnisgewinnung im Chemieunterricht der Sekundarstufe I. *Entwicklung und quantitative Dimensionalitätsanalyse eines Testinstruments aus epistemologischer Perspektive*
ISBN 978-3-8325-5256-5 44.00 EUR

312 Christina Kobl: Förderung und Erfassung der Reflexionskompetenz im Fach Chemie
ISBN 978-3-8325-5259-6 41.00 EUR

313 Ann-Kathrin Beretz: Diagnostische Prozesse von Studierenden des Lehramts – *eine Videostudie in den Fächern Physik und Mathematik*
ISBN 978-3-8325-5288-6 45.00 EUR

314 Judith Breuer: Implementierung fachdidaktischer Innovationen durch das Angebot materialgestützter Unterrichtskonzeptionen. *Fallanalysen zum Nutzungsverhalten von Lehrkräften am Beispiel des Münchener Lehrgangs zur Quantenmechanik*
ISBN 978-3-8325-5293-0 50.50 EUR

315 Michaela Oettle: Modellierung des Fachwissens von Lehrkräften in der Teilchenphysik. *Eine Delphi-Studie*
ISBN 978-3-8325-5305-0 57.50 EUR

316 Volker Brüggemann: Entwicklung und Pilotierung eines adaptiven Multistage-Tests zur Kompetenzerfassung im Bereich naturwissenschaftlichen Denkens
ISBN 978-3-8325-5331-9 40.00 EUR

317 Stefan Müller: Die Vorläufigkeit und soziokulturelle Eingebundenheit naturwissenschaftlicher Erkenntnisse. *Kritische Reflexion, empirische Befunde und fachdidaktische Konsequenzen für die Chemielehrer*innenbildung*
ISBN 978-3-8325-5343-2 63.00 EUR

318 Laurence Müller: Alltagsentscheidungen für den Chemieunterricht erkennen und Entscheidungsprozesse explorativ begleiten
ISBN 978-3-8325-5379-1 59.00 EUR

319 Lars Ehlert: Entwicklung und Evaluation einer Lehrkräftefortbildung zur Planung von selbstgesteuerten Experimenten
ISBN 978-3-8325-5393-71 41.50 EUR

320 Florian Seiler: Entwicklung und Evaluation eines Seminarkonzepts zur Förderung der experimentellen Planungskompetenz von Lehramtsstudierenden im Fach Chemie
ISBN 978-3-8325-5397-5 47.50 EUR

321 Nadine Boele: Entwicklung eines Messinstruments zur Erfassung der professionellen Unterrichtswahrnehmung von (angehenden) Chemielehrkräften hinsichtlich der Lernunterstützung
ISBN 978-3-8325-5402-6 46.50 EUR

322 Franziska Zimmermann: Entwicklung und Evaluation digitalisierungsbezogener Kompetenzen von angehenden Chemielehrkräften
ISBN 978-3-8325-5410-1 49.50 EUR

323 Lars-Frederik Weiß: Der Flipped Classroom in der Physik-Lehre. *Empirische Untersuchungen in Schule und Hochschule*
ISBN 978-3-8325-5418-7 51.00 EUR

324 Tilmann Steinmetz: Kumulatives Lehren und Lernen im Lehramtsstudium Physik. *Theorie und Evaluation eines Lehrkonzepts*
ISBN 978-3-8325-5421-7 51.00 EUR

325 Kübra Nur Celik: Entwicklung von chemischem Fachwissen in der Sekundarstufe I. *Validierung einer Learning Progression für die Basiskonzepte „Struktur der Materie“, „Chemische Reaktion“ und „Energie“ im Kompetenzbereich „Umgang mit Fachwissen“*
ISBN 978-3-8325-5431-6 55.00 EUR

326 Matthias Ungermann: Förderung des Verständnisses von Nature of Science und der experimentellen Kompetenz im Schüler*innen-Labor Physik in Abgrenzung zum Regelunterricht
ISBN 978-3-8325-5442-2 55.50 EUR

327 Christoph Hoyer: Multimedial unterstütztes Experimentieren im webbasierten Labor zur Messung, Visualisierung und Analyse des Feldes eines Permanentmagneten
ISBN 978-3-8325-5453-8 45.00 EUR

328 Tobias Schüttler: Schülerlabore als interessefördernde authentische Lernorte für den naturwissenschaftlichen Unterricht nutzen
ISBN 978-3-8325-5454-5 50.50 EUR

329 Christopher Kurth: Die Kompetenz von Studierenden, Schülerschwierigkeiten beim eigenständigen Experimentieren zu diagnostizieren
ISBN 978-3-8325-5457-6 58.50 EUR

330 Dagmar Michna: Inklusiver Anfangsunterricht Chemie *Entwicklung und Evaluation einer Unterrichtseinheit zur Einführung der chemischen Reaktion*
ISBN 978-3-8325-5463-7 49.50 EUR

331 Marco Seiter: Die Bedeutung der Elementarisierung für den Erfolg von Mechanikunterricht in der Sekundarstufe I
ISBN 978-3-8325-5471-2 66.00 EUR

332 Jörn Hägele: Kompetenzaufbau zum experimentbezogenen Denken und Arbeiten. *Videobasierte Analysen zu Aktivitäten und Vorstellungen von Schülerinnen und Schülern der gymnasialen Oberstufe bei der Bearbeitung von fachmethodischer Instruktion*
ISBN 978-3-8325-5476-7 56.50 EUR

333 Erik Heine: Wissenschaftliche Kontroversen im Physikunterricht. *Explorationsstudie zum Umgang von Physiklehrkräften und Physiklehramtsstudierenden mit einer wissenschaftlichen Kontroverse am Beispiel der Masse in der Speziellen Relativitätstheorie*
ISBN 978-3-8325-5478-1 48.50 EUR

334 Simon Goertz: Module und Lernzirkel der Plattform FLexKom zur Förderung experimenteller Kompetenzen in der Schulpraxis *Verlauf und Ergebnisse einer Design-Based Research Studie*
ISBN 978-3-8325-5494-1 66.50 EUR

335 Christina Toschka: Lernen mit Modellexperimenten *Empirische Untersuchung der Wahrnehmung und des Denkens in Analogien beim Umgang mit Modellexperimenten*
ISBN 978-3-8325-5495-8 50.00 EUR

336 Alina Behrendt: Chemiebezogene Kompetenzen in der Übergangsphase zwischen dem Sachunterricht der Primarstufe und dem Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-5498-9 40.50 EUR

337 Manuel Daiber: Entwicklung eines Lehrkonzepts für eine elementare Quantenmechanik *Formuliert mit In-Out Symbolen*
ISBN 978-3-8325-5507-8 48.50 EUR

338 Felix Pawlak: Das Gemeinsame Experimentieren (an-)leiten *Eine qualitative Studie zum chemiespezifischen Classroom-Management*
ISBN 978-3-8325-5508-5 46.50 EUR

339 Liza Dopatka: Konzeption und Evaluation eines kontextstrukturierten Unterrichtskonzeptes für den Anfangs-Elektrizitätslehreunterricht
ISBN 978-3-8325-5514-6 69.50 EUR

340 Arne Bewersdorff: Untersuchung der Effektivität zweier Fortbildungsformate zum Experimentieren mit dem Fokus auf das Unterrichtshandeln
ISBN 978-3-8325-5522-1 39.00 EUR

341 Thomas Christoph Münster: Wie diagnostizieren Studierende des Lehramtes physikbezogene Lernprozesse von Schüler*innen? Eine Videostudie zur Mechanik
ISBN 978-3-8325-5534-4 44.50 EUR

342 Ines Komor: Förderung des symbolisch-mathematischen Modellverständnisses in der Physikalischen Chemie
ISBN 978-3-8325-5546-7 46.50 EUR

343 Verena Petermann: Überzeugungen von Lehrkräften zum Lehren und Lernen von Fachinhalten und Fachmethoden und deren Beziehung zu unterrichtsnahem Handeln
ISBN 978-3-8325-5545-0 47.00 EUR

344 Jana Heinze: Einfluss der sprachlichen Konzeption auf die Einschätzung der Qualität instruktionaler Unterrichtserklärungen im Fach Physik
ISBN 978-3-8325-5545-0 47.00 EUR

345 Jannis Weber: Mathematische Modellbildung und Videoanalyse zum Lernen der Newtonschen Dynamik im Vergleich
ISBN 978-3-8325-5566-5 68.00 EUR

346 Fabian Sterzing: Zur Lernwirksamkeit von Erklärvideos in der Physik *Eine Untersuchung in Abhängigkeit von ihrer fachdidaktischen Qualität und ihrem Einbettungsformat*
ISBN 978-3-8325-5576-4 52.00 EUR

347 Lars Greitemann: Wirkung des Tablet-Einsatzes im Chemieunterricht der Sekundarstufe I unter besonderer Berücksichtigung von Wissensvermittlung und Wissenssicherung
ISBN 978-3-8325-5580-1 50.00 EUR

348 Fabian Poensgen: Diagnose experimenteller Kompetenzen in der laborpraktischen Chemielehrer*innenbildung
ISBN 978-3-8325-5587-0 48.00 EUR

349 William Lindlahr: Virtual-Reality-Experimente *Entwicklung und Evaluation eines Konzepts für den forschend-entwickelnden Physikunterricht mit digitalen Medien*
ISBN 978-3-8325-5595-5 49.00 EUR

350 Bert Schlüter: Teilnahmemotivation und situationales Interesse von Kindern und Eltern im experimentellen Lernsetting KEMIE
ISBN 978-3-8325-5598-6 43.00 EUR

351 Katharina Nave: Charakterisierung situativer mentaler Modellkomponenten in der Chemie und die Bildung von Hypothesen *Eine qualitative Studie zur Operationalisierung mentaler Modell-komponenten für den Fachbereich Chemie*
ISBN 978-3-8325-5599-3 43.00 EUR

352 Anna B. Bauer: Experimentelle Kompetenz Physikstudierender *Entwicklung und erste Erprobung eines performanzorientierten Kompetenzstrukturmodells unter Nutzung qualitativer Methoden*
ISBN 978-3-8325-5625-9 47.00 EUR

353 Jan Schröder: Entwicklung eines Performanztests zur Messung der Fähigkeit zur Unterrichtsplanung bei Lehramtsstudierenden im Fach Physik
ISBN 978-3-8325-5655-9 46.50 EUR

354 Susanne Gerlach: Aspekte einer Fachdidaktik Körperpflege *Ein Beitrag zur Standardentwicklung*
ISBN 978-3-8325-5659-4 45.00 EUR

355 Livia Murer: Diagnose experimenteller Kompetenzen beim praktisch-naturwissenschaftlichen Arbeiten *Vergleich verschiedener Methoden und kognitive Validierung eines Testverfahrens*
ISBN 978-3-8325-5657-0 41.50 EUR

356 Andrea Maria Schmid: Authentische Kontexte für MINT-Lernumgebungen *Eine zweiteilige Interventionsstudie in den Fachdidaktiken Physik und Technik*
ISBN 978-3-8325-5605-1 57.00 EUR

357 Julia Ortmann: Bedeutung und Förderung von Kompetenzen zum naturwissenschaftlichen Denken und Arbeiten in universitären Praktika
ISBN 978-3-8325-5670-9 37.00 EUR

358 Axel-Thilo Prokop: Entwicklung eines Lehr-Lern-Labors zum Thema Radioaktivität *Eine didaktische Rekonstruktion*
ISBN 978-3-8325-5671-6 49.50 EUR

359 Timo Hackemann: Textverständlichkeit sprachlich variierter physikbezogener Sachtexte
ISBN 978-3-8325-5675-4 41.50 EUR

360 Dennis Dietz: Vernetztes Lernen im fächerdifferenzierten und integrierten naturwissenschaftlichen Unterricht aufgezeigt am Basiskonzept Energie *Eine Studie zur Analyse der Wirksamkeit der Konzeption und Implementation eines schulinternen Curriculums für das Unterrichtsfach „Integrierte Naturwissenschaften 7/8“*
ISBN 978-3-8325-5676-1 49.50 EUR

361 Ann-Katrin Krebs: Vielfalt im Physikunterricht *Zur Wirkung von Lehrkräftefortbildungen unter Diversitätsaspekten*
ISBN 978-3-8325-5672-3 65.50 EUR

362 Simon Kaulhausen: Strukturelle Ursachen für Klausurmisserfolg in Allgemeiner Chemie an der Universität
ISBN 978-3-8325-5699-0 37.50 EUR

Studien zum Physik- und Chemielernen

Herausgegeben von Martin Hopf und Mathias Ropohl

Die Reihe umfasst inzwischen eine große Zahl von wissenschaftlichen Arbeiten aus vielen Arbeitsgruppen der Physik- und Chemiedidaktik und zeichnet damit ein gültiges Bild der empirischen physik- und chemiedidaktischen Forschung im deutschsprachigen Raum.
Die Herausgeber laden daher Interessenten zu neuen Beiträgen ein und bitten sie, sich im Bedarfsfall an den Logos-Verlag oder an ein Mitglied des Herausgeberteams zu wenden.

Kontaktadressen:

Univ.-Prof. Dr. Martin Hopf
Universität Wien,
Österreichisches Kompetenzzentrum
für Didaktik der Physik,
Porzellangasse 4, Stiege 2,
1090 Wien, Österreich,
Tel. +43-1-4277-60330,
e-mail: martin.hopf@univie.ac.at

Prof. Dr. Mathias Ropohl
Didaktik der Chemie,
Fakultät für Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen,
Tel. 0201-183 2704,
e-mail: mathias.ropohl@uni-due.de

Um experimentelle Kompetenzen gezielt zu fördern, bedarf es einer differenzierten Diagnose, z. B. durch Tests mit Realexperimenten. Von Schüler:innen angefertigte Protokolle bieten dabei eine zeitökonomische Möglichkeit, die Kompetenzen zu erfassen. Es stellt sich aber die Frage, ob sie eine genaue Diagnose zulassen und inwiefern zusätzliche Methoden, wie Videos oder Interviews, die Genauigkeit der Diagnose erhöhen. In der vorliegenden Studie wurden die Ergebnisse der Diagnose bei Tests mit Realexperimenten anhand verschiedener Methoden verglichen. Zudem wurde das Testverfahren kognitiv validiert.

Hierfür bearbeiteten 27 Jugendliche jeweils vier Experimentieraufgaben zum naturwissenschaftlichen Messen ($N = 108$). Währenddessen füllten sie Protokolle aus und wurden videografiert sowie anschließend interviewt. Um Hinweise auf kognitive Validität zu untersuchen, wurden die Aussagen der Schüler:innen in den Interviews ausgewertet und von Expert:innen hinsichtlich der Passung mit den durch die Aufgaben intendierten Konzepten zum naturwissenschaftlichen Messen eingeschätzt.

Die Ergebnisse zeigen, dass für eine genaue Diagnose experimenteller Kompetenzen bei Tests mit Realexperimenten Protokolle und Interviews nötig sind. Zudem ließen sich Hinweise für kognitive Validität finden. Somit konnte eine kognitiv valide Möglichkeit zur Diagnose experimenteller Kompetenzen im Bereich des Messens aufgezeigt werden, die nahelegt, Schüler:innen auch zu ihrem Handeln beim Experimentieren zu befragen.

Logos Verlag Berlin

ISBN 978-3-8325-5657-0