Knowledge of what to assess in Bezug auf Experimentierkompetenzen – Modellierung, Messung und Validierung einer Facette von *assessment literacy* von Lehramtsstudierenden

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen

im Promotionsprogramm Biologie
der Georg-August University School of Science (GAUSS)

vorgelegt von
Cora Joachim
aus Salzgitter

Göttingen, 2020

Betreuungsausschuss

Prof. Dr. Susanne Bögeholz, Didaktik der Biologie, Georg-August-Universität Göttingen

Prof. Dr. Marcus Hammann, Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster

Mitglieder der Prüfungskommission

Referentin: Prof. Dr. Susanne Bögeholz, Didaktik der Biologie, Georg-August-Universität Göttingen

Korreferent: Prof. Dr. Marcus Hammann, Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Dieter Heineke, Fakultät für Biologie und Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Susanne Schneider, Didaktik der Physik, Georg-August-Universität Göttingen

Prof. Dr. Sascha Schroeder, Pädagogische Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Thomas Waitz, Fachdidaktik Chemie, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 12.06.2020

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

1D	one-dimensional
3D	three-dimensional
AAAS	American Association for the Advancement of Science
AIC	Akaike Information Criterion
ANOVA	analysis of variance
BIC	Bayesian Information Criterion
BMBF	Bundesministerium für Bildung und Forschung
CK	content knowledge
DIF	differential item functioning
EAP/PV	expected a posteriori, based on plausible values
ExMo	Verbund-Projekt *Vermittlungs- und Beurteilungskompetenzen zum Experimentieren: Modellierung, Validierung und Messinstrumententwicklung*
GFD	Gesellschaft für Fachdidaktik
IRT	item response theory
KMK	Kultusministerkonferenz
KoKoHs	Förderinitiative *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor*
NGSS	Next Generation Science Standards
NRC	National Research Council
PCK	pedagogical content knowledge
PISA	Programme for International Student Assessment
PK	pedagogical knowledge
POSITT	Pedagogy of Science Inquiry Teaching Test
SDDS	Scientific Discovery as Dual Search
TH	Testheft
Tukey HSD test	Tukey honestly significant difference test
wMNSQ	weighted mean square

Zusammenfassung und Abstract

Zusammenfassung

Assessment literacy von Lehrkräften ist von großer Bedeutung für das Lernen von Schüler*innen. Lehrkräfte benötigen sowohl Fähigkeiten im Unterrichten als auch im Beurteilen, um ihre Schüler*innen optimal fördern zu können. Im Biologieunterricht stellen Kompetenzen im Experimentieren ein zentrales Lernziel für Schüler*innen dar. Die Komplexität dieser Kompetenzen erfordert eine schrittweise Vermittlung. Lehrkräften kommt die Aufgabe zu, die Kompetenzen fortlaufend zu beurteilen, um den Unterricht bestmöglich an die Lernvoraussetzungen der Schüler*innen zu adaptieren. Dafür müssen sie über Wissen zu Kriterien des Experimentierens und möglichen Vorstellungen von Schüler*innen zum Experimentieren verfügen und dieses Wissen anwenden können (*knowledge of what to assess*).

Bislang liegen kaum Erkenntnisse zum Wissen für die Beurteilung von Experimentierkompetenzen vor. Das Ziel der vorliegenden Dissertation und den damit verbundenen Studien ist die Modellierung, Messung und Validierung von *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Experimentierkompetenzen. Damit einher geht die Entwicklung eines Messinstruments.

Die Auswertungen zeigen, dass es möglich ist, ein Instrument zur Erfassung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen mit offenen Aufgaben zu entwickeln. Hinweise auf Validität liegen vor. Die in der Hauptstudie (N=500) vorgenommene Modellierung mit Item-Response-Theorie Modellen (IRT) gibt Einblick in die Dimensionalität des Konstrukts. Eine Analyse der Qualität der Beurteilungen zeigt Stärken und Schwächen von Biologielehramtsstudierenden in der Beurteilung von Experimentierkompetenzen. Unter anderem stellen die Beurteilung von Vorstellungen von Schüler*innen zum Experimentieren und die Verwendung von Fachbegriffen in der Beurteilung eine Herausforderung für Biologielehramtsstudierende dar. Damit gibt die Arbeit Hinweise auf weiteres Potenzial für die Förderung von Biologielehramtsstudierenden in Bezug auf *assessment literacy*.

Abstract

Teachers' assessment literacy is important for students' learning. Teachers need both skills in teaching and assessing to foster their students in an optimum way. A core learning goal in biology is experimentation competences. These complex competences have to be taught successively. A continuous assessment of students' experimentation competences by their teachers is important to adapt the instruction to students' prerequisites. For that purpose, teachers have to know criteria for experimentation as well as potential student conceptions regarding experimentation and be able to apply this knowledge.

Up to now little is known about teachers' knowledge regarding the assessment of experimentation competences. This dissertation and the connected studies aim to model, measure and validate knowledge of what to assess of pre-service biology teachers of experimentation competences. This involves the development of a measurement instrument.

The analyses demonstrate that it is possible to develop a measurement instrument for knowledge of what to assess of experimentation competences applying an open answer format. Evidence for validity is given. The modeling conducted in the main study (N=500) with item response theory models (IRT) informs about dimensionality of the construct. An analysis of the quality of the assessments demonstrates strengths and weaknesses of pre-service teachers in assessing experimentation competences. For instance, it is a challenge for pre-service teachers to assess student conceptions and to apply technical terms in the assessments. Therewith, the study informs teacher education about further potential in fostering pre-service teachers' assessment literacy.

1. Einleitung

Beurteilungsfähigkeiten von Lehrkräften wird eine große Bedeutung für den Lernfortschritt von Schüler*innen zugeschrieben (Brunner et al., 2011; Ruiz-Primo & Furtak, 2007). Unterricht ist besonders lernwirksam, wenn er auf die Lernvoraussetzungen, z.B. Schwierigkeiten, der Schüler*innen abgestimmt ist (Beck et al., 2008; Weinert et al., 1990, S. 169). Beurteilungen ermöglichen es Lehrkräften, den Lernstand und die Lernfortschritte ihrer Schüler*innen in Bezug auf die Lernziele zu erfassen und bilden damit eine Grundlage für die Anpassung des Unterrichts an die Lernvoraussetzungen der Schüler*innen (Donovan & Bransford, 2005, S. 16; Weinert, 2000). Besonders zentral sind formative Beurteilungen, die Unterrichtsentscheidungen leiten und z.B. dazu dienen, den Lernvoraussetzungen der Schüler*innen entsprechende Aufgaben und Aktivitäten zu wählen (Praetorius & Südkamp, 2017, S. 13; Schrader, 2006, S. 95; 2008, S. 169; Weinert et al., 1990, S. 169). Informelle formative Beurteilungen, die, im Gegensatz zu formellen Beurteilungen (z.B. Zensurengebung), laut Schrader (2006) „eher beiläufig und unsystematisch“ (S. 95) vorgenommen werden und Unterrichtsentscheidungen dienen, können häufig in einer Unterrichtsstunde erfolgen und gehören damit zu den Kernaufgaben von Lehrkräften (Black et al., 2003, S. 2; Schrader & Helmke, 1990, S. 312f.). Wissen für Beurteilungen stellt einen Aspekt des Professionswissens von Lehrkräften dar (Baumert & Kunter, 2006; Magnusson, et al., 1999). Baumert und Kunter (2006, S. 489) fassen diagnostische Kompetenz vornehmlich als allgemeines pädagogisches Wissen und Können, vermuten aber auch eine Abhängigkeit von fachdidaktischem Wissen. Magnusson et al. (1999) ordnen Wissen für Beurteilungen dem fachdidaktischen Wissen (*pedagogical content knowledge*) zu und betonen damit die Fachspezifität. Nach Magnusson et al. (1999) umfasst *knowledge of assessment of scientific literacy*[1] Wissen zu *dimensions of science learning to assess* (z.B. *nature of science*, *scientific investigation* und *practical reasoning*) und *methods of assessing science learning* (z.B. schriftliche Tests und praktische Untersuchungen) (S. 108f.).

Die Bedeutung der Ausbildung von Lehrkräften im Beurteilen wurde erkannt. Der Erwerb von Kompetenzen zur Beurteilung von Leistungen von Schüler*innen ist national und international ein Ziel der Lehrerbildung (Kultusministerkonferenz, 2019a, 2019b; National Research Council, 1996). Die ländergemeinsamen inhaltlichen Anforderungen für

[1] Für eine internationale Anbindung werden die englischen Bezeichnungen verwendet. Dies gilt ebenso für das Konstrukt *assessment literacy* und dessen Komponenten.

die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung geben vor, dass angehende Lehrkräfte im Studium u.a. Wissen über „Grundlagen fach- bzw. fachrichtungs- und anforderungsgerechter Leistungsbeurteilung“ erwerben sollen (Kultusministerkonferenz, 2019b, S. 4). Am Ende des Vorbereitungsdienstes sollen Lehrkräfte die „fach- bzw. fachrichtungsspezifische Leistungsbeurteilung beherrschen“ (Kultusministerkonferenz, 2019b, S. 4).

Ein fachspezifischer Blick auf Biologie unterstreicht die Bedeutung formativer Beurteilungen: Die komplexen Kompetenzen in der Biologie müssen schrittweise über einen längeren Zeitraum vermittelt werden (Töpperwien & Köttker, 2010, S. 4). Entsprechend wichtig für ihre Förderung ist die fortlaufende Beurteilung von Lernzwischenständen. Lehrkräfte sollten u.a. in der Lage sein, Kompetenzen der Schüler*innen beim Experimentieren – einer zentralen Methode der Erkenntnisgewinnung der Biologie, die sehr anspruchsvoll für Schüler*innen ist und spezifisch gefördert werden muss – zu beurteilen (Hammann et al., 2006; Klautke, 1997, S. 323; Kultusministerkonferenz, 2005; Schulz et al., 2012, S. 15). Für die Beurteilung von Experimentierkompetenzen ist Wissen über Anforderungen an das Experimentieren (Kriterien des Experimentierens[2]) und Vorstellungen der Schüler*innen zum Experimentieren grundlegend (vgl. Kapitel 5, Bögeholz, Joachim[3] et al., 2016 und Kapitel 6, Joachim et al., 2020).

Studien zum Wissen für die Beurteilung von Schüler*innenkompetenzen in den Naturwissenschaften sind begrenzt (vgl. Kapitel 6, Joachim et al., 2020). Eine Prüfung des Forschungsstands zum Wissen von Lehrkräften der Naturwissenschaften ergab, dass nur wenige Studien vorliegen, die Wissen von Lehrkräften für die Beurteilung direkt erfassen (Abell, 2007, S. 1131f.; Abell & Siegel, 2011, S. 207). Eine Studie im deutschsprachigen Raum untersuchte diagnostische Kompetenzen angehender Biologielehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung (Dübbelde, 2013). Die eingesetzten Aufgaben erfassten, inwiefern vorgegebene Anforderungen an das Experimentieren, wie z.B. Bezug der Schlussfolgerung zur Hypothese, zur Beurteilung angewandt werden konnten, nicht jedoch inwiefern Studierende über dieses Wissen für Beurteilungen (u.a. Wissen zu den Anforderungen an das Experimentieren und Vorstellungen von Schüler*innen) selbst verfügen. Der Cronbachs alpha des Instruments (17 Items) lag bei 0.50 (Dübbelde, 2013, S. 189).

[2] Hier wird Bezug auf Kriterien genommen, damit die Zuordnung über Rahmen und Publikationen hinweg möglich ist. In Publikation 3 entspricht das *criteria*.

[3] gemeinsame Erstautorenschaft

Die vorliegende Arbeit ergänzt die Forschung zum Wissen von Lehrkräften für die Beurteilung von naturwissenschaftlich relevanten Experimentierkompetenzen von Schüler*innen und dessen Anwendung. Die Arbeit ist eingebettet in das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Verbund-Projekt *Vermittlungs- und Beurteilungskompetenzen zum **Ex**perimentieren: **Mo**dellierung, Validierung und Messinstrumentenentwicklung* (ExMo)[4]. Am ExMo Projekt waren neben der Georg-August-Universität Göttingen (Prof. Dr. S. Bögeholz, C. Joachim) die Westfälische Wilhelms-Universität Münster (Prof. Dr. M. Hammann, S. Hasse) und Otto-Friedrich-Universität Bamberg (Prof. Dr. C. H. Carstensen) beteiligt. In zwei Teilprojekten wurden Kompetenzen zur Vermittlung von Experimentierkompetenzen (Westfälische Wilhelms-Universität Münster) und Wissen und Fähigkeiten für Beurteilungen (Georg-August-Universität Göttingen) untersucht (Bögeholz et al., 2013). Die Projektbearbeitung fand zeitgleich und inhaltlich und methodisch aufeinander abgestimmt statt. Der Fokus dieser Arbeit liegt auf dem Göttinger Teilprojekt. Ziel ist die Modellierung, Messung und Validierung von *knowledge of what to assess* (Abell & Siegel, 2011) – einem für Beurteilungen grundlegenden Wissensbereich – in Bezug auf Experimentierkompetenzen.

Die im Rahmen der Arbeit entstandenen Publikationen stellen die im Forschungsprozess gewonnenen Erkenntnisse dar. Sie fokussieren 1) die Entwicklung von Messinstrumenten im ExMo Projekt (Kapitel 4, Hasse et al., 2014), 2) Hinweise einer Pilotierung auf die Reliabilität und Validität im Göttinger Teilprojekt zur Beurteilungskompetenz (Kapitel 5, Bögeholz, Joachim et al., 2016) und 3) die Modellierung, Messung und Validierung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen (Kapitel 6, Joachim et al., 2020). Die Publikationen werden im Folgenden vom theoretischen Hintergrund (Kapitel 2) und den daraus abgeleiteten Forschungsfragen (Kapitel 3) sowie einer Zusammenfassung und Diskussion (Kapitel 7) eingerahmt.

[4] Das Göttinger ExMo Teilprojekt wurde im Rahmen der Förderinitiative *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor* (KoKoHs) unter dem Förderkennzeichen 01PK11014B vom BMBF gefördert.

2. Theoretischer Hintergrund und empirische Befunde

Im Folgenden wird der theoretische Hintergrund zu Wissen und Fähigkeiten[5] von (angehenden) Lehrkräften für die Beurteilung von Experimentierkompetenzen von Schüler*innen dargestellt. Zunächst werden Konzeptualisierungen von Beurteilungskompetenz beleuchtet und verschiedene Ziele und Gegenstände von Beurteilungen im Unterricht ausgeführt (Kapitel 2.1) sowie für die Arbeit relevante Modelle zu *assessment literacy* vorgestellt (Kapitel 2.2). Im dritten Abschnitt werden Erkenntnisse zu den zu beurteilenden Experimentierkompetenzen von Schüler*innen dargelegt (Kapitel 2.3). Darauf aufbauend werden in Abschnitt vier Erkenntnisse aus den Kapiteln 2.2 und 2.3 zusammengeführt und *knowledge of what to assess* in Bezug auf Experimentierkompetenzen von Schüler*innen in Biologie dargestellt (Kapitel 2.4). Schließlich werden die Konstrukte *Beurteilungskompetenz* und *knowledge of what to assess* voneinander abgegrenzt (Kapitel 2.5).

2.1 Beurteilungen von Schüler*innen zur Förderung adaptiven Unterrichtens

Adaptiver Unterricht beruht auf der differenzierten Anpassung des Unterrichts an die Lernvoraussetzungen der Schüler*innen, sodass individuelles Lernen gefördert wird (Beck et al., 2008, S. 47; Fischer, et al., 2014; Wember & Melle, 2018, S. 58). Erforderlich dafür ist die Fähigkeit der Lehrkräfte, die Lernvoraussetzungen der Schüler*innen zu beurteilen (Fischer et al., 2014).

Beurteilungskompetenz bzw. Diagnosekompetenz[6] wird u.a. als Urteilsgenauigkeit von Lehrkräften in Bezug auf Personenmerkmale definiert und operationalisiert (Schrader, 2006, S. 95; Schrader & Helmke, 1987; Spinath, 2005). Die Urteilsgenauigkeit wird in Form der Übereinstimmung der Urteile der Lehrkräfte mit den bei Schüler*innen gemessenen Merkmalen untersucht (Schrader, 2013, S. 157). Diese Operationalisierung grenzt Diagnosekompetenz stark ein. Eine Definition von Weinert hingegen drückt ein breiteres Verständnis aus. Nach Weinert (2000) handelt es sich bei diagnostischen Kompetenzen „um ein Bündel von Fähigkeiten, um den Kenntnisstand, die Lernfortschritte und die Leis-

[5] Entsprechend der Herangehensweise der OECD im Rahmen der PISA-Studien werden Wissen und Fähigkeiten (Englisch: *knowledge and skills*), als Teilaspekte von Kompetenz verstanden (OECD, 2005a, 2005b). Teilweise wird der Begriff „Fertigkeiten“ gleichbedeutend zu oder in Kombination mit „Fähigkeiten“ verwendet (Emden, 2011, S. 12; OECD, 2005b).
Nach Klieme et al. (2004, S. 70) entwickelt sich Wissen zu Fähigkeiten mit zunehmender Kompetenz. Wir begreifen Fähigkeiten nicht als wissensunabhängig (Méhaut & Winch, 2012, S. 374).

[6] Beurteilungskompetenz und Diagnosekompetenz werden oft gleichgestellt (Schrader, 2009, S. 237).

tungsprobleme der einzelnen Schüler sowie die Schwierigkeiten verschiedener Lernaufgaben im Unterricht fortlaufend beurteilen zu können, sodass das didaktische Handeln auf diagnostischen Einsichten aufgebaut werden kann“ (S. 14). Auch Schrader (2011), der Diagnosekompetenz als Urteilsgenauigkeit definiert und operationalisiert hat, erweitert die Definition und fasst diagnostische Kompetenz als die „Gesamtheit der zur Bewältigung von Diagnoseaufgaben erforderlichen Fähigkeiten“ (S. 683).

Diagnosen sind vielfältig. Sie können u.a. in Bezug auf ihr Ziel und den diagnostizierten Gegenstand unterschieden werden (Aufschnaiter et al., 2015). Mit Blick auf das Ziel von Beurteilungen unterscheiden Black et al. (2003, S. 1f.) formelle Beurteilungen von informellen. Formelle Beurteilungen dienen dazu, bestimmte Bildungsabschlüsse zu bescheinigen und Schulen anhand von Leistungsergebnissen zu vergleichen, um damit Schulen in die Verantwortung zu nehmen. Beurteilungen hingegen, die dem Lernen dienen, sind üblicherweise informell und Teil des regulären Unterrichts (ebd.). Damit diese informellen Beurteilungen lernförderlich sind, müssen sie Informationen liefern, die der verbesserten Abstimmung des Unterrichts auf die Lernbedürfnisse der Schüler*innen dienen. Bei dieser Form von Beurteilungen handelt es sich um formative Beurteilungen, die von den Lehrkräften häufig – unabhängig von dem Zweck der Leistungsüberprüfung – vorgenommen werden (ebd., S. 2).

In Bezug auf den diagnostizierten Gegenstand unterscheiden Aufschnaiter et al. (2015, S. 744-747) vier Arten von Diagnosen: 1) Statusdiagnostik fokussiert einen Lernstand (Feststellung einer vorliegenden Kompetenz oder eines Merkmals), 2) Prozessdiagnostik betrachtet „die zu einem Zeitpunkt stattfindenden Prozesse des Handelns und Denkens“ (ebd., S. 245), 3) Veränderungsdiagnostik umfasst den Vergleich von mindestens zwei Status- oder Prozessdiagnosen und ermittelt, inwiefern eine Veränderung stattgefunden hat und 4) Verlaufsdiagnostik hat die Art und Weise der Entwicklung von Kompetenzen über einen längeren Zeitraum zum Gegenstand. Die Diagnostikarten zwei bis vier haben die Betrachtung von Lernprozessen gemein. Statusdiagnostik kann Teil von prozessbezogenen Diagnosen sein und ist die Basis für adaptiven Unterricht (ebd., S. 739, 744-747). Aufgrund der Breite des Konstrukts, das z.B. die Fähigkeiten umfasst, fachspezifische Vorstellungen von Schüler*innen zu identifizieren, Diagnoseverfahren sinnvoll auszuwählen und anzuwenden und Diagnoseaufgaben zu entwickeln, wird eine Mehrdimensionalität des Konstrukts Diagnosekompetenz angenommen (ebd., S. 739).

Nach Dübbelde et al. (2010) erfordert fachbezogene Diagnosekompetenz Fachwissen (*content knowledge*, CK), fachdidaktisches Wissen (*pedagogical content knowledge*,

PCK) und pädagogisches Wissen (*pedagogical knowledge*, PK). Die Vorgaben für die Lehrerbildung sehen den Erwerb von Wissen primär im Studium vor. Auf dieses Wissen wird im Vorbereitungsdienst aufgebaut. Dieser praktische Ausbildungsabschnitt soll die Entwicklung von Kompetenzen, z.B. in der Planung und Gestaltung von fachlichem Lernen oder der fachlichen Leistungsbeurteilung auf Basis des Wissens, ermöglichen (Kultusministerkonferenz 2019b, S. 4). Die „Quantität und Qualität der Lerngelegenheiten" (Voss, 2019, S. 14) im Studium erweisen sich als entscheidend für den Erwerb des Wissens (ebd.; Kunter et al., 2011, S. 57). So konnten auch beispielsweise signifikante Unterschiede im CK (Wissen zu Evolution, Genetik, Mikrobiologie und Morphologie) und PCK (u.a. Wissen zum Verständnis von Schüler*innen und Wissen für Beurteilungen) von Biologielehramtsstudierenden an deutschen Universitäten mit unterschiedlich vielen Lerngelegenheiten festgestellt werden (Großschedl et al., 2015). Studierende, die im Studium weiter fortgeschritten waren, erreichten höhere Testergebnisse für CK und PCK als Studienanfänger*innen (Großschedl et al., 2015). Auch in Bezug auf Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung konnte ein Anstieg für Fach- und Lehramtsstudierende der Physik vom Bachelorstudium zum Masterstudium ermittelt werden (Straube, 2016).

Im internationalen Raum werden Wissen und Fähigkeiten von Lehrkräften für Beurteilungen unter dem Begriff *assessment literacy* gefasst (Abell & Siegel, 2011, S. 206). Im Folgenden werden zwei für unsere Studie besonders relevante Modelle zu *assessment literacy* vorgestellt.

2.2 Modelle zu *assessment literacy* von Lehrkräften

Basierend auf Magnusson et al. (1999) und Pellegrino et al. (2001) präsentieren Abell und Siegel (2011) ein Modell zu *assessment literacy*, das vier Bereiche von Wissen unterscheidet (siehe Abbildung 2-1). Im Zentrum ihres Modells steht die Vorstellung vom Lernen. Damit verknüpft sind Werte und Prinzipien (Abell & Siegel, 2011, S. 211). Versteht eine Lehrkraft beispielsweise Lernen als Konstruktion von Wissen und legt Wert auf das Prinzip, dass Beurteilungsaufgaben dem Lernen dienen, setzt sie ggf. problemorientierte, offene Aufgaben anstatt von Multiple-Choice Fragen zur Beurteilung ein (Abell & Siegel, 2011, S. 212). Beurteilungen basieren auf der Anwendung von Wissen, das von Abell und Siegel (2011) in vier Bereiche unterteilt wird, die sich in der Praxis gegenseitig beeinflussen. Die vier Bereiche sind: 1) *knowledge of assessment purposes,* 2) *knowledge of what to assess,* 3) *knowledge of assessment strategies* und

4) *knowledge of assessment interpretation and action-taking* (Abell & Siegel, 2011, S. 213ff.). Sie werden im Folgenden beschrieben.

1) *knowledge of assessment purposes:* Abell und Siegel (2011) unterscheiden vier Arten von Beurteilungen in Bezug auf das Beurteilungsziel. Dazu zählen a) diagnostische Beurteilungen, die zu Beginn einer Einheit vorgenommen werden, Informationen zum Wissensstand und den Vorstellungen der Schüler*innen liefern und eine Orientierung für den Unterricht geben, b) formative Beurteilungen, die während des Unterrichts erfolgen, eine Rückmeldefunktion für Lehrkräfte und Schüler*innen haben und helfen, den Unterricht auf die Bedürfnisse der Schüler*innen zu adaptieren, c) summative Beurteilungen, die der Dokumentation des Lernerfolgs z.B. am Ende einer Einheit dienen und oftmals eine Grundlage von Noten darstellen und d) metakognitive Beurteilungen, die Schüler*innen helfen, ihr eigenes Lernen zu überprüfen und zu steuern. Metakognitive Beurteilungen können mit den unter a) bis c) genannten Beurteilungsarten zusammen erfolgen (Abell & Siegel, 2011, S. 213f.).

2) *Knowledge of what to assess* steht in Bezug zu curricularen Vorgaben und entsprechenden Lernzielen. Lehrkräfte müssen vertraut sein mit den Kompetenzen, die im Kerncurriculum verankert sind und von Schüler*innen erworben werden sollen, um zu beurteilen, inwiefern diese Kompetenzen erreicht werden. Auch ist es erforderlich, typische Fehlvorstellungen bei Schüler*innen erkennen zu können, um letztendlich auch adaptiv unterrichtlich darauf reagieren zu können. Dabei beeinflussen Werte der Lehrkraft, welche Lernziele sie als besonders wichtig erachtet und entsprechend in Beurteilungen fokussiert (Abell & Siegel, 2011, S. 214, 217).

3) *Knowledge of assessment strategies* fasst Wissen über verschiedene Strategien, wie beispielsweise Prüfungsaufgaben für formelle Beurteilungen und Tests für informelle Beurteilungen. Zudem ordnen Abell und Siegel diesem Wissensbereich Wissen über Strategien zur Rückmeldung zu. Dazu gehören verschiedene Formen von Feedback, wie z.B. Selbst- und Peer-Bewertung, und Methoden, die den Schüler*innen helfen, das Feedback zu nutzen (Abell & Siegel, 2011, S. 214f.).

4) *Knowledge of assessment interpretation and action-taking* beinhaltet Wissen über die Verwendung der in Beurteilungen gewonnenen Erkenntnisse. Beispielsweise können Beurteilungsergebnisse für die Notengebung oder die Adaptation des Unterrichts genutzt werden (Abell & Siegel, 2011, S. 215).

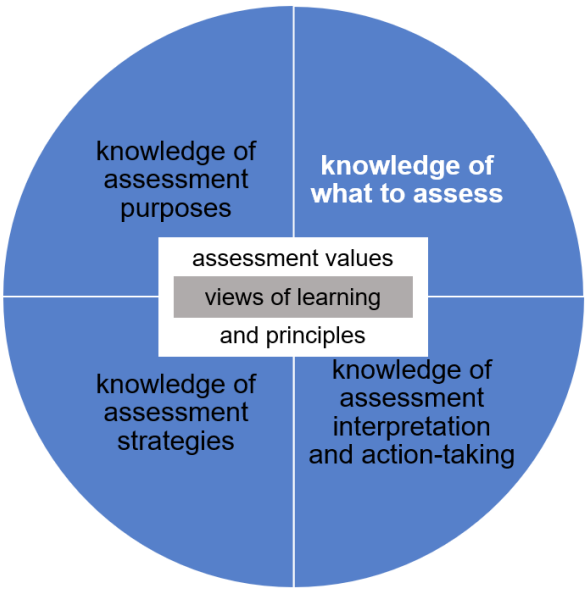


Abbildung 2-1 Modell zu assessment literacy von Lehrkräften der Naturwissenschaften (verändert nach Abell & Siegel, 2011, S. 212)

Das zweite Modell zu *assessment literacy*, das dieser Arbeit zugrunde liegt, stammt von Xu und Brown (2016). Ihr Modell betont ebenfalls die Bedeutung von Wissen für Beurteilungen und gliedert dieses noch weiter auf als das Modell von Abell und Siegel (2011) (siehe Abbildung 2-2 und Abbildung 2-3). Auch die von Abell und Siegel (2011) im Zentrum lokalisierten Werte und Prinzipien in Bezug auf Beurteilungen greift das Modell von Xu und Brown (2016) in Form von Vorstellungen auf. Zusätzlich berücksichtigt das Modell von Xu und Brown (2016) Faktoren die *assessment literacy* beeinflussen und stellt die Ausprägung von *assessment literacy* hierarchisch in Form einer Pyramide dar, beginnend mit dem Wissen über das *was*, *warum* und *wie* von Beurteilungen als Fundament und der Identifizierung als Beurteiler*in als der am weitesten entwickelten Stufe von *assessment literacy* an der Spitze der Pyramide (Xu & Brown, 2016, S. 154f., 159). Im Folgenden werden die Komponenten des Modells im Einzelnen vorgestellt.

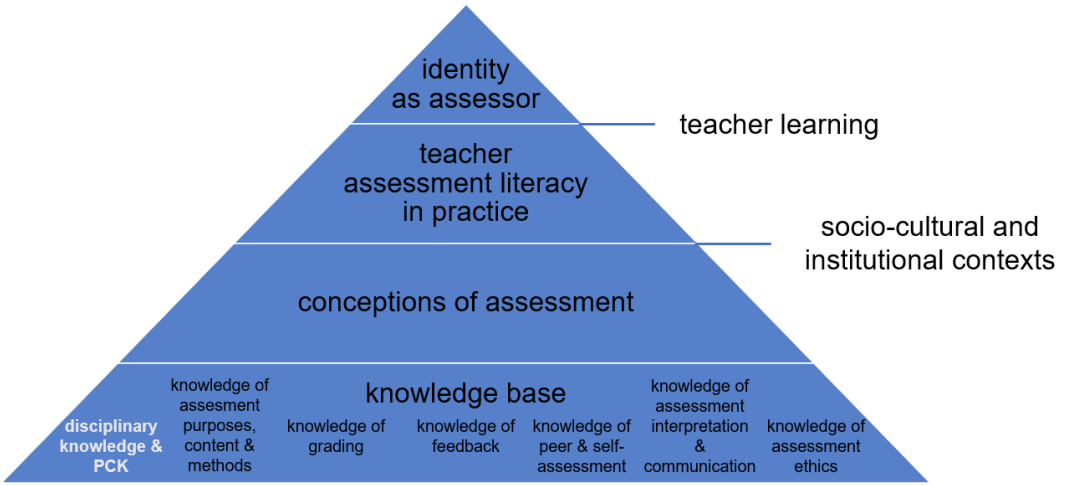


Abbildung 2-2 Modell zu assessment literacy in der Praxis (verändert nach Xu & Brown, 2016, S. 155)

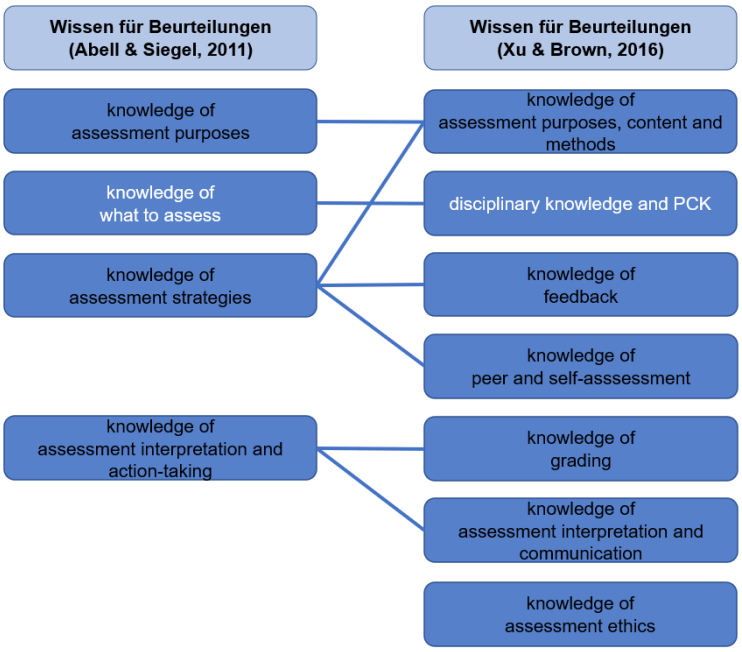


Abbildung 2-3 Wissensbereiche von assessment literacy nach Abell & Siegel (2011) und Xu & Brown (2016)

1) Die unterste Stufe der Pyramide bildet die Wissensbasis. Die Wissensbasis ist die Grundlage aller anderen Komponenten und umfasst sieben Wissensbereiche: a) *disciplinary knowledge and PCK*, b) *knowledge of assessment purposes, content and methods,* c) *knowledge of grading,* d) *knowledge of feedback,* e) *knowledge of peer and self-assessment,* f) *knowledge of assessment interpretation and communication* und

g) *knowledge of assessment ethics.* Der Bereich a) *disciplinary knowledge and PCK* (Xu & Brown, 2016) entspricht 2) *knowledge of what to assess* (Abell & Siegel, 2011) (siehe Abbildung 2-3). Xu und Brown (2016) begründen die Relevanz dieses Wissensbereichs folgendermaßen: „Since educational assessment is about measuring the curriculum content taught in schools/universities, knowledge of disciplines and how to teach that content cannot be excluded from the assessment knowledge base" (S. 156). Der Bereich b) *knowledge of assessment purposes, content and methods* umfasst Wissen, das Abell und Siegel (2011) in die beiden Bereiche 1) *knowledge of assessment purposes* und 3) *knowledge of assessment strategies* unterteilen. Den Wissensbereich 3) *knowledge of assessment strategies* (Abell & Siegel, 2011) unterteilen Xu und Brown (2016) weiterhin in d) *knowledge of feedback* und e) *knowledge of peer and self-assessment.* Als weitere Bereiche der Wissensbasis führen Xu und Brown (2016) c) *knowledge of grading* und f) *knowledge of assessment interpretation and communication* auf. Beide zusammen sind inhaltlich vergleichbar mit dem Wissensbereich 4) *knowledge of assess-ment interpretation and action-taking* von Abell und Siegel (2011), umfassen jedoch nicht das Ableiten von Folgerungen für den Unterricht. Zudem enthält die Wissensbasis von Xu und Brown (2016) den Bereich g) *knowledge of assessment ethics*, z.B. in Bezug auf die Verwendung, Aufbewahrung und Weitergabe von Beurteilungsergebnissen sowie die Gerechtigkeit von Beurteilungen allen Lernenden gegenüber (S. 156). Dieser Aspekt findet sich bei Abell und Siegel (2011) unter Werten und Prinzipien, die Lehrkräfte bei Beurteilungen anlegen.

2) Die zweitunterste Stufe der Pyramide bilden Vorstellungen der Lehrkräfte von Beurteilungen, die auf Basis von Vorerfahrungen entstanden sein können. Die Vorstellungen beeinflussen, inwiefern neues Wissen für Beurteilungen aufgenommen und umgesetzt wird (Xu & Brown, 2016, S. 156).

3) Über den Vorstellungen der Lehrkräfte stehen soziokulturelle und institutionelle Kontexte. Damit wird berücksichtigt, dass Lehrkräfte in ihren Beurteilungen an gewisse Vorgaben, z.B. Standards, gebunden sind, die sich je nach Arbeitsumfeld unterscheiden können. Neben offiziellen Vorgaben, die für Beurteilungen bestehen, können auch Schuladministrator*innen, Kolleg*innen sowie Eltern und Schüler*innen Beurteilungen beeinflussen, indem sie mit bestimmten Bedürfnissen und Interessen an die Lehrkräfte herantreten (Xu & Brown, 2016, S. 157).

4) An nächst höherer Stelle in der Pyramide nennen Xu und Brown (2016) *teacher assessment literacy in practice* (TALiP). TALiP legt einen Fokus auf die praktische Durchführung von Beurteilungen. Xu und Brown (2016) verdeutlichen damit, dass Beurteilungen äußeren Einflüssen unterliegen und Lehrkräfte beim Beurteilen Kompromisse zwischen äußeren Faktoren, wie beispielsweise offiziellen Vorgaben oder Interessen von Kolleg*innen und Eltern etc., und eigenen Vorstellungen, z.B. in Bezug auf das Ziel von Beurteilungen, eingehen müssen (S. 156f.). TALiP ist nach Xu and Brown (2016) „a dynamic, complex entity combining teachers' assessment knowledge, their conceptions of assessment, and their responses to the external contexts embedded with actual constraints and affordances in the environment" (S. 157).

5) Für die Verbesserung von TALiP ist Lehrerbildung zentral. Als ein Ziel in der Lehrerbildung sehen Xu und Brown (2016), bei Lehrkräften ein Hinterfragen von ggf. vorliegenden Vorstellungen in Bezug auf Beurteilungen herbeizuführen, um eine Weiterentwicklung von *assessment literacy* in Gang zu setzen (S. 157). Reflexionen und der Austausch mit Kolleg*innen können das Lernen über Beurteilungen und damit die Verbesserung von TALiP ermöglichen (ebd., S. 158).

6) Als höchstes Ziel sehen Xu und Brown (2016), dass Lehrkräfte neben dem klassischen Rollenverständnis ein Verständnis von sich als Beurteiler*innen vom Lernen haben. Das bedeutet, dass Lehrkräfte die Bedeutung von Beurteilungen verstehen und Beurteilungen als ihre Aufgabe begreifen und diese entsprechend planen, durchführen und auswerten (S. 158).

Sowohl das Modell von Abell und Siegel (2011) als auch von Xu und Brown (2016) heben die Bedeutung verschiedener Wissensbereiche als Grundlage für Beurteilungen hervor. In der vorliegenden Arbeit, die Wissen für die Beurteilung von Experimentierkompetenzen im Biologieunterricht untersucht, wird ein besonders fachspezifischer Teilbereich dieser Wissensbasis fokussiert: *Knowledge of what to assess* bzw. *disciplinary knowledge and PCK* in Bezug auf Experimentierkompetenzen in Biologie.

2.3 Experimentierkompetenzen von Schüler*innen

Erkenntnisgewinnung ist einer von vier Kompetenzbereichen, die in den nationalen Bildungsstandards der Bundesrepublik Deutschland für das Fach Biologie verankert sind (Kultusministerkonferenz, 2005). Experimentieren stellt/e in der Vergangenheit, Gegenwart und Zukunft eine ganz zentrale Methode der Erkenntnisgewinnung in Biologie dar (Klautke, 1997, S. 323; Schulz et al., 2012, S. 15). Die Bedeutung des Experimentierens für Schüler*innen spiegelt sich sowohl national wie international durch die Verankerung

in Standards des Kompetenzbereichs Erkenntnisgewinnung bzw. in den *National Science Education Standards* wider (Kultusministerkonferenz, 2005; National Research Council, 1996). In Bezug auf das Experimentieren sollen Schüler*innen mit Abschluss von Klasse 10 in der Lage sein, „einfache Experimente“ zu planen, durchzuführen und auszuwerten (E6) sowie „Schritte aus dem experimentellen Weg der Erkenntnisgewinnung zur Erklärung“ anzuwenden (E7) (Kultusministerkonferenz, 2005, S. 14).

Das „Experimentieren [kann] als ein komplexer Prozess des Problemlösens aufgefasst werden“ (Hammann, 2004, S. 198; Klahr, 2000). Verschiedene Modellierungen naturwissenschaftlicher Erkenntnisprozesse umfassen drei bis fünf „Prozessvariablen“ (Wellnitz et al., 2012, S. 264), wie z.B. die „Suche im Hypothesenraum“, das „Testen von Hypothesen“ und die „Analyse von Evidenzen“ (Hammann et al., 2007, S. 44). Zwei für die Naturwissenschaftsdidaktiken besonders relevante Modelle, das *Scientific Discovery as Dual Search*-Modell (SDDS-Modell, Klahr, 2000) und das Strukturmodell zum wissenschaftlichen Denken (Mayer, 2007; Mayer et al., 2008), werden im Folgenden vorgestellt.

Grundlage des SDDS-Modells ist das Verständnis von Erkenntnisgewinnung als Problemlöseprozess. Das SDDS-Modell besteht aus drei Komponenten: Der Prozess der Erkenntnisgewinnung beginnt mit der „Suche im Hypothesenraum“ und der damit verbundenen Bildung spezifischer Hypothesen zur Erklärung eines Phänomens (Hammann, 2007, S. 189; Hammann et al., 2007, S. 35; Klahr, 2000, S. 29f.). Als zweite Komponente folgt das „Testen von Hypothesen“ mit einem aussagekräftigen Experiment (umfasst u.a. die Suche im Experiment-Raum) (Hammann, 2007, S. 189; Hammann et al., 2007, S. 35; Klahr, 2000, S. 30, 35). Zum Schluss wird die „Analyse von Evidenzen“ vorgenommen und entschieden, ob die untersuchte Hypothese angenommen, widerlegt oder weiter untersucht wird (Hammann et al., 2007, S. 35f.; Klahr, 2000, S. 30ff.; Klautke, 1997, S. 324; Li & Klahr, 2006). Das SDDS-Modell mit seinen drei Komponenten wurde im Modell zu Experimentierkompetenzen aufgegriffen (Hammann, 2004; Hammann et al., 2007). In einer Untersuchung der Dimensionalität von Experimentierkompetenzen von Fünft- und Sechstklässlern mit Multiple-Choice Aufgaben konnten zwei Dimensionen festgestellt werden. Testen von Hypothesen konnte als eine Dimension des Experimentierens identifiziert werden. Die Suche im Hypothesenraum und Analyse von Evidenzen wurden zusammen als zweite Dimension eingestuft. Bedingt sein könnte dies ggf. durch die stärkere Abhängigkeit der Aufgaben zur Suche im Hypothesenraum und Analyse von Evidenzen von inhaltlichem Vorwissen als von methodischem Vorwissen. Im Gegensatz dazu ist für das Lösen der Aufgaben zum Testen von Hypothesen hauptsächlich methodisches Wissen relevant (Hammann et al., 2007, S. 43, 45).

Im Modell zum wissenschaftlichen Denken (Mayer, 2007) werden vier Prozessvariablen unterschieden: „Naturwissenschaftliche Fragen formulieren“, „Hypothesen generieren“, „Untersuchungen planen“ und „Daten analysieren/Schlussfolgerungen ziehen“ (S. 181). Vier Teilkompetenzen, die die Bewältigung der mit den vier Prozessvariablen verbundenen Anforderungen fokussieren, konnte Grube (2010, S. 56ff.) in ihrer Modellierung empirisch fundieren. Für die vier Teilkompetenzen konnten signifikante Leistungsunterschiede bei Schüler*innen aus Klasse 5-10 gezeigt werden (ebd., S. 62f.). Die Deutung der Ergebnisse fällt Schüler*innen am leichtesten. Das Generieren von Hypothesen fällt ihnen etwas schwerer als die Deutung, aber leichter als die Planung einer Untersuchung und insbesondere das Formulieren von naturwissenschaftlichen Fragen (ebd., S. 63). Die mit Zehntklässlern durchgeführte Studie von Wellnitz (2012) zur Struktur der Kompetenz „Naturwissenschaftliche Untersuchungen“, die ebenfalls die vier Phasen „Fragestellung“, „Hypothese“, „Untersuchungsdesign“ und „Datenauswertung“ umfasst, deutet auf eine Eindimensionalität des Konstrukts hin (S. 131f.). Aber auch Wellnitz (2012) stellt unterschiedliche Personenfähigkeiten in den vier Phasen fest (abhängig von den für die Berechnung verwendeten Aufgaben sind die Phasen mit der höchsten bis zur niedrigsten Personenfähigkeit: Datenauswertung, Fragestellung, Hypothese, Untersuchungsdesign bzw. unter Kontrolle der Antwortformate, Komplexitätsniveaus und kognitiven Prozesse: Untersuchungsdesign, Datenauswertung, Fragestellung, Hypothese) (S. 139ff.). Jeder einzelne Schritt im Prozess der Erkenntnisgewinnung von der Fragestellung bis zur Datenauswertung hat bestimmte Anforderungen, die Schwierigkeiten für Schüler*innen darstellen können (Hammann et al., 2006). Diese Anforderungen werden im Folgenden zusammengefasst (vgl. dazu auch Kapitel 5, Bögeholz, Joachim et al., 2016 und Kapitel 6, Joachim et al., 2020).

Experimente dienen der Untersuchung kausaler Zusammenhänge (Klautke, 1997, S. 323; Wellnitz, 2012, S. 31). „Bei Experimenten schließt die Fragestellung das kausale Verhältnis (Ursache – Wirkung) zwischen unabhängiger und abhängiger Variable ein“ (Bruckermann et al., 2017, S. 16). Naturwissenschaftliche Fragestellungen sollen sich auf ein spezifisches Phänomen beziehen und naturwissenschaftlich überprüfbar sein (ebd., S. 17; Marbach-Ad & Claassen, 2001, S. 418). Das Generieren naturwissenschaftlicher Fragestellungen erfordert Hintergrundwissen über das jeweilige Thema (Marbach-Ad & Claassen, 2001, S. 418).

Ein wichtiges Kriterium für Hypothesen ist die Testbarkeit (Klautke, 1997, S. 324; Li & Klahr, 2006; Mayer & Ziemek, 2006, S. 6). Dafür werden Hypothesen präzise formuliert

und die vermutete Beziehung zwischen unabhängiger und abhängiger Variable angegeben (Mayer & Ziemek, 2006, S. 6; Wellnitz, 2012, S. 45f.). Zudem sollen Hypothesen anhand von Vorwissen „theoretisch begründet werden“ (Mayer & Ziemek, 2006, S. 6). Die Suche im Hypothesenraum soll umfassend erfolgen, d.h. alle möglichen Erklärungen für ein Phänomen müssen berücksichtigt werden (Hammann et al., 2006, S. 298; Wellnitz, 2012, S. 36). Das kann besonders für jüngere Schüler*innen schwierig sein. Ihnen gelingt es oft nicht, Hypothesen zu generieren, die nicht den eigenen Vorstellungen entsprechen (Hammann et al., 2006, S. 298; Klahr et al., 1993, S. 125, 137; Koslowski, 1996, S. 245; vgl. Kapitel 6, Joachim et al., 2020).

Das Testen von Hypothesen erfordert einen systematischen Umgang mit Variablen. Experimental- und Kontrollansatz dürfen sich nur in der Ausprägung einer unabhängigen Variable unterscheiden (Hammann et al., 2006, S. 292f.; Klautke, 1997, S. 323; Schulz et al., 2012, S. 18f.). Die abhängige Variable wird beobachtet und gemessen (Krüger, 2009, S. 43; Mayer & Ziemek, 2006, S. 7). Schüler*innen variieren die Variablen oft unsystematisch und/oder berücksichtigen keine Kontrollansätze (Chen & Klahr, 1999; Hammann et al., 2006, S. 292ff.). Häufig ergeben sich Fehler aus der Vorstellung, das Ziel des Experimentierens sei es, einen Effekt zu erzielen (bezeichnet als „Ingenieursmodus“ (Hammann & Mayer, 2012, S. 284)) (Li & Klahr, 2006, S. 12; Schauble et al., 1991, S. 860ff.). Schüler*innen legen dann keinen Fokus auf die Untersuchung von „Ursache-Wirkungs-Beziehungen“ (Hammann & Mayer, 2012, S. 284; Schauble et al., 1991, S. 860ff.). Experimentieren erfordert eine genaue Planung und die Kontrolle aller Bedingungen. Ein Experiment muss objektiv durchgeführt und ausgewertet werden. Das Vorgehen und die Daten werden dokumentiert, damit das Experiment wiederholbar ist (Krüger, 2009, S. 43; Mayer & Ziemek, 2006, S. 7). Die Relevanz von Messwiederholungen ist vielen Schüler*innen nicht bewusst (Duggan & Gott, 2000, S. 207; Lubben & Millar, 1996, S. 958f.; vgl. Kapitel 6, Joachim et al., 2020).

In der Analyse von Evidenzen werden die Daten präzise ausgewertet und Fehler diskutiert (Mayer & Ziemek, 2006, S. 6f.). Die Ergebnisse werden auf die Hypothesen bezogen (ebd., S. 6). Die Hypothesen werden entsprechend der Ergebnisse gestützt, widerlegt oder weiter untersucht (Hammann et al., 2007, S. 35f.; Klahr, 2000, S. 30ff.; Klautke, 1997, S. 324; Li & Klahr, 2006). Schüler*innen kann es schwerfallen, Hypothesen zu widerlegen, die ihrem Vorwissen entsprechen (Li & Klahr, 2006, S. 9). Ggf. kann ein „Bestätigungsbias“ (Hammann et al., 2006, S. 297) dazu führen, dass den Erwartungen widersprechende Ergebnisse ignoriert oder z.B. methodischen Fehlern zugeschrieben werden (Chinn & Brewer, 1998, S. 628; Li & Klahr, 2006, S. 9f.). Die Hypothesen werden

ggf. trotz widerlegender Daten nicht verworfen (Chinn & Brewer, 1993; vgl. Kapitel 6, Joachim et al., 2020).

2.4 *Knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Experimentierkompetenzen

Auf Basis des theoretischen Hintergrunds zu Wissen und Fähigkeiten in Bezug auf Beurteilungen (Kapitel 2.2) und zu Experimentierkompetenzen (Kapitel 2.3) wird das Konstrukt *knowledge of what to assess* in Bezug auf Experimentierkompetenzen von Schüler*innen dargestellt.

Das für Lehrkräfte notwendige Wissen für die Beurteilung von Experimentierkompetenzen umfasst Wissen über die Lernziele für Schüler*innen. Lehrkräfte sollten in der Lage sein, zu erfassen, inwiefern Lernziele erreicht werden und Schülerfehlvorstellungen vorliegen (Abell & Siegel, 2011, S. 217), um den Unterricht lernförderlicher gestalten zu können. Ein gezieltes Eingehen auf Schülervorstellungen, u.a. das Auslösen von Unzufriedenheit mit den vorliegenden Vorstellungen, ist wichtig, um eine Veränderung dieser Vorstellungen zu ermöglichen (Krüger, 2007). Entsprechend sollten Lehrkräfte zentrale Kriterien der Hypothesenbildung, Planung von Experimenten und Auswertung von Daten (vgl. Kapitel 2.3) kennen und die Schüler*innenlösungen in Bezug darauf beurteilen können.[7] Z.B. sollten Lehrkräfte in der Lage sein, zu beurteilen, inwiefern Hypothesen von Schüler*innen testbar und begründet sind und inwiefern die Suche im Hypothesenraum von Schüler*innen umfassend erfolgt. Zudem ist es hilfreich für Lehrkräfte, mit den häufig vorkommenden Schülervorstellungen beim Experimentieren (Ingenieursmodus, *Confirmation Bias*/Bestätigungsbias) vertraut zu sein und diese identifizieren und erklären zu können (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016 und Kapitel 6, Joachim et al., 2020).

Es liegen nur wenige Studien zum Wissen von Naturwissenschaftslehrkräften für Beurteilungen vor (Abell & Siegel, 2011, S. 207). Beispielsweise untersuchte Cappell (2013) *Diagnosekompetenz* angehender Physiklehrkräfte. Hinsichtlich der Beurteilung von Schülervorstellungen stellte sie fest, dass angehenden Physiklehrkräften das Nennen

[7] Unter Berücksichtigung des SDDS-Modells bzw. des Modells zu Experimentierkompetenzen nach Hammann (2004) und der Überschneidung von Anforderungen an Fragestellungen und Hypothesen (naturwissenschaftlich überprüfbar, Bezug zur unabhängigen und abhängigen Variable, Beachtung von Hintergrundwissen) (vgl. Kapitel 2.3), umfasst die vorliegende Modellierung von *knowledge of what to assess* Kriterien des Experimentierens in Bezug auf die drei Phasen *Hypothesenbildung*, *Planung von Experimenten* und *Auswertung von Daten*.

von Vorstellungen von Schüler*innen leichter fällt als das Identifizieren von Fehlvorstellungen in Vignetten (S. 166f.). Sie fand Hinweise darauf, dass Studierende sowohl in der Nennung als auch der Identifikation von Fehlvorstellungen im Studienverlauf besser werden. Außerdem sind angehende Physiklehrkräfte in der Lage, Fehlvorstellungen im eigenen Unterricht zu erkennen (Cappell, 2013, S. 167f.).

In Bezug auf die fachspezifische Beurteilung von Experimentierkompetenzen in Biologie wurde eine Studie von Dübbelde (2013) durchgeführt. In der Studie kam u.a. ein Instrument zur Erfassung der statusdiagnostischen Kompetenz von angehenden Biologielehrkräften zum Einsatz. Den angehenden Lehrkräften lag ein Lösungsheft zweier Schüler zu einem Experiment sowie ein Beurteilungsbogen dazu vor. Die Studierenden mussten ankreuzen, inwiefern die Schüler bestimmte Anforderungen an das Experiment erfüllt hatten. Ein Item lautete beispielsweise: „Versuchsplanung berücksichtigt Messwiederholungen“ (Dübbelde, 2013, S. 25 im Anhang). Verwendet wurden die folgenden Antwortkategorien: „Ja“, „Nein“, „weiß nicht“ (Dübbelde, 2013, S. 25 im Anhang). Das Testformat erfasst, inwiefern Biologielehramtsstudierende bestimmte Kriterien zum Experimentieren in Schüler*innenlösungen identifizieren können. Offen bleibt jedoch, inwiefern die Studierenden selbst diese Kriterien, die für Experimente zentral sind, kennen und zur Beurteilung anwenden können. Zudem führt das Testformat mit zumeist drei Antwortalternativen, eine davon „weiß nicht“, zu einer hohen Ratewahrscheinlichkeit. Die Reliabilität des Instruments wurde mit einem Cronbachs alpha Wert von 0.50 angegeben.

Neben *Diagnosekompetenz* zeigt das Konstrukt *Untersuchen* (Krüger et al., 2014; Straube, 2016) Überschneidungen mit *knowledge of what to assess* in Bezug auf Experimentierkompetenzen. Die von Krüger et al. (2014, S. 2) für Studierende entwickelte Kurzskala zum Untersuchen in der Biologie umfasst, ähnlich wie Experimentierkompetenzen, die Phasen „Fragen formulieren“, „Hypothesen generieren“, „Planen von Untersuchungen“ und „Auswertung von Untersuchungen“ im Kontext von Biologie. Wissen zu diesen Phasen ist Teil von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen.

Weiterhin wird ein Zusammenhang von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen mit entsprechenden Selbstwirksamkeitserwartungen angenommen (vgl. Kapitel 6, Joachim et al., 2020). Der Fragebogen von Mahler (2014) erfasst Selbstwirksamkeitserwartungen u.a. hinsichtlich der Planung von Biologieunterricht unter Berücksichtigung von biologiedidaktischen Forschungsergebnissen (z.B. Erkenntnisse zu Schülervorstellungen und Kompetenzen im Fach Biologie) sowie Kompetenzen

und Basiskonzepten. Darin inbegriffen sind Erkenntnisse zu Schülervorstellungen und Experimentierkompetenzen – beides gehört zum Wissen für die Beurteilung von Experimentierkompetenzen. Hohe Selbstwirksamkeitserwartungen können die Anwendung des Wissens verbessern (Bouffard-Bouchard et al., 1991; Tschannan-Moran et al., 1998, S. 211).

Ein Zuwachs von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen im Verlauf des Studiums wird angenommen (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016 und Kapitel 6, Joachim et al., 2020), da Wissen für Beurteilungen, Kenntnisse und Fertigkeiten im Experimentieren und Wissen über Schülervorstellungen in der Biologie Inhalte der universitären Lehrerbildung sein sollten (Kultusministerkonferenz, 2019b).

Das Konstrukt *knowledge of what to assess* in Bezug auf Experimentierkompetenzen ist vergleichsweise eng ausgerichtet, indem es nur einen Wissensbereich von *assessment literacy* umfasst. Nichtsdestotrotz ist, z.B. vor dem Hintergrund, dass die Beurteilung von Experimentierkompetenzen Wissen zum Experimentieren voraussetzt, eine mehrdimensionale Struktur möglich. Studien zu Kompetenzen von Schüler*innen im Bereich naturwissenschaftlicher Erkenntnisgewinnung kommen zu unterschiedlichen Ergebnissen in Bezug auf die Dimensionalität der Kompetenzen: Während die Studie von Wellnitz zur Struktur der Kompetenz *naturwissenschaftliche Untersuchungen* (Wellnitz, 2012, S. 131) auf Eindimensionalität hindeutet, sprechen die Ergebnisse von Grube (2010, S. 56ff.) und Hammann et al. (2007, S. 42) für eine Mehrdimensionalität der Konstrukte *wissenschaftliches Denken* und *Experimentierkompetenzen* (vgl. Kapitel 2.3).

2.5 Konstrukt Beurteilungskompetenz bzw. *knowledge of what to assess*

Die Bezeichnung des untersuchten Konstrukts wurde aufgrund eines tiefergehenden Verständnisses im Projektverlauf präzisiert. In der deutschsprachigen Publikation (Kapitel 5, Bögeholz, Joachim et al., 2016) wurde *knowledge of what to assess* unter der Kurzbezeichnung *Beurteilungskompetenz* verhandelt und in der ersten, konzeptionellen Publikation (Kapitel 4, Hasse et al., 2014) wurde zunächst vorläufig (weniger fokussiert) von *assessment competence* gesprochen.

Bei Antrag des ExMo Projekts war die Untersuchung von Beurteilungskompetenz vorgesehen (vgl. Kapitel 4, Hasse et al., 2014 und Kapitel 5, Bögeholz, Joachim et al., 2016). Im Verlauf des Projekts wurde die Komplexität von Beurteilungskompetenz (vgl. Kapitel 2.1) erarbeitet. Dazu zählt auch die Definition von diagnostischer Kompetenz als die „Gesamtheit der zur Bewältigung von Diagnoseaufgaben erforderlichen Fähigkei-

ten“ (Schrader, 2011, S. 683). Unser Ansatz bei der Modellierung, Messung und Validierung von Beurteilungskompetenz operationalisiert diese hingegen als wissensbasierten Aspekt von Beurteilungen. Folglich wurden Beurteilungsaufgaben für Biologielehramtsstudierende entwickelt, die testen, inwiefern (angehende) Biologielehrkräfte über Wissen für die Beurteilung von Experimentierkompetenzen ihrer Schüler*innen verfügen und dabei die Anforderungen an kompetentes Experimentieren anlegen und bei der Beurteilung als Kriterien anwenden (vgl. Kapitel 6, Joachim et al., 2020). Das gemessene Konstrukt ist damit im Vergleich zu diagnostischer Kompetenz (vgl. Schrader, 2011; Weinert, 2000) weniger breit.

Der in dieser Arbeit verfolgte Ansatz berücksichtigt die Bedeutung von Wissen als grundlegende Voraussetzung für Beurteilungskompetenz bzw. *assessment literacy* (Xu & Brown, 2016, S. 159) und steht damit in Einklang mit den Vorgaben für die Lehrerbildung. Diese sehen vor, dass Biologielehramtsstudierende im Studium zunächst grundlegendes Wissen für Beurteilungen erwerben (Kultusministerkonferenz, 2019b, S. 4). Erst mit dem Vorbereitungsdienst wird erwartet, dass die Leistungsbeurteilung beherrscht wird (Kultusministerkonferenz, 2019b, S. 4). Entsprechend ist von Studierenden nicht zu erwarten, dass sie bereits über Beurteilungskompetenz im Sinne von umfassenden Fähigkeiten für vielfältige Beurteilungsaufgaben verfügen. Mit Blick auf die Kompetenzentwicklung von angehenden Biologielehrkräften in der universitären Lehrerbildung widmet sich die vorliegende Arbeit wissensbasierten Aspekten von Beurteilungen mit dem Untersuchungsschwerpunkt *knowledge of what to assess* in Bezug auf Experimentierkompetenzen (vgl. Kapitel 6, Joachim et al., 2020). Bislang liegen wenige Erkenntnisse zu *knowledge of what to assess* vor, die sich auf Experimentierkompetenzen beziehen. Die Entwicklung eines Messinstruments ist zentral für die Erfassung von *knowledge of what to assess*, um Bildungsprozesse zu evaluieren. Die drei Publikationen, die im Rahmen der Arbeit entstanden sind, widmen sich bei der Messinstrumentenentwicklung sowie der Modellierung und Validierung dem gleichen Konstrukt, das mit *knowledge of what to assess* in Bezug auf Experimentierkompetenzen am prägnantesten bezeichnet wird (vgl. Kapitel 6, Joachim et al., 2020).

3. Forschungsfragen und ihre Bearbeitung

Die Betrachtung der Modelle zu *assessment literacy* (Kapitel 2.2) hat gezeigt, dass Beurteilungen eine breite Wissensbasis voraussetzen, die sich von *knowledge of assessment purposes* über *knowledge of what to assess* bis hin zu *knowledge of assessment interpretation and action-taking* erstreckt. Bei *knowledge of what to assess* (Abell & Siegel, 2011) handelt es sich, vergleichbar mit *disciplinary knowledge and PCK* (Xu & Brown, 2016), um einen besonders fachbezogenen Bereich der Wissensbasis, der in Bezug auf das Fach Biologie z.B. Wissen über Lernziele und Vorstellungen zum Experimentieren voraussetzt. Für die einzelnen Phasen des Experimentierens bestehen spezifische Anforderungen, die Schüler*innen Schwierigkeiten bereiten können und entsprechend von Lehrkräften in der Beurteilung der Experimentierkompetenzen berücksichtigt werden müssen (Kapitel 2.3 und Kapitel 2.4). Der Blick auf den Forschungsstand zeigt, dass noch weiterer Forschungsbedarf in der Untersuchung des Wissens für die Beurteilung von Experimentierkompetenzen besteht (Kapitel 2.4). Die vorliegende Arbeit ergänzt die Forschung zu *knowledge of what to assess* in Bezug auf Experimentierkompetenzen mittels eines offenen, realitätsnahen Aufgabenformats, das die Anwendung von Wissen erfordert. Die Aufgaben entsprechen informellen Beurteilungen von Lernständen (Statusdiagnostik) (vgl. Kapitel 2.1).

Das Ziel der Arbeit ist daher die Modellierung, Messung und Validierung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen. Ein zentraler Schritt dafür ist zunächst die theoriegeleitete Entwicklung eines Messinstruments zur Erfassung des Konstrukts. Die Messinstrumententwicklung umfasst mehrere Schritte und hat eine reliable und valide Messung zum Ziel. Damit können Erkenntnisse z.B. zur Dimensionalität des Konstrukts sowie zu Stärken und Schwächen der Lehramtsstudierenden hinsichtlich *knowledge of what to assess* in Bezug auf Experimentierkompetenzen gewonnen werden.

Daraus ergeben sich folgende Forschungsfragen:

(1) Inwiefern kann ein Instrument zur reliablen Messung des Konstrukts *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Schülerexperimentierkompetenzen entwickelt werden? (Publikationen 1, 2, 3)
(2) Inwiefern gibt es Hinweise auf Ein- oder Mehrdimensionalität des Konstrukts? (Publikation 3)
(3) Inwiefern gibt es Hinweise auf Validität? (Publikationen 1, 2, 3)

(4) Welche Stärken und Schwächen haben Biologielehramtsstudierende in der Beurteilung von Experimentierkompetenzen von Schüler*innen? (Publikation 3)

Das Messinstrument wurde sukzessive von Publikation 1 (Kapitel 4, Hasse et al., 2014) über Publikation 2 (Kapitel 5, Bögeholz, Joachim et al., 2016) hin zu Publikation 3 (Kapitel 6, Joachim et al., 2020) entwickelt. Die Publikation 3 enthält schließlich die Modellierung, die Messung und die Validierung. Erkenntnisse zu Stärken und Schwächen von Biologielehramtsstudierenden hinsichtlich *knowledge of what to assess* in Bezug auf Experimentierkompetenzen werden dargestellt und Potenzial zur weiteren Förderung des Wissens und der Fähigkeiten in der universitären Ausbildung in der Didaktik der Biologie wird aufgezeigt.

4. Assessing Teaching and Assessment Competences of Biology Teacher Trainees: Lessons from Item Development[8]

Abstract

In Germany, science education standards for students at the end of grade nine have been in existance since 2005. Some of these standards are dedicated to scientific inquiry (e.g. experimentation). They describe which abilities learners are expected to possess at the end of grade nine. In the USA, several documents describe standards for *Teaching Inquiry* (NGSS 2013, NRC 1996/2000/2007, AAAS 1989). Presently, comparable teaching standards for science teachers are mostly lacking in Germany. Further, there are hardly any instruments that allow for the assessment of specific competences pertaining to teaching experimental lessons and assessing student competences in experimentation. Therefore, the aim of the project described in this paper is to develop assessment instruments for biology teachers who are being trained at universities as well as in in-service teacher training programs with respect to i) analyzing experimental biology lessons, ii) planning experimental biology lessons, and iii) assessing student achievements in experimental biology lessons. The article gives insights into ongoing research with respect to assessing the quality of biology teacher education. Finally, the developed measurement instruments should allow for assessing the learning preconditions of future biology teachers. The instruments offer first starting points for the development of sensitive measures for longitudinal studies to investigate university teacher education and teacher traineeship in the subject of biology.

Key words: Science education, Biology teacher trainees, Measurement instrument, Pedagogical content knowledge, Experimentation.

4.1 Introduction

The concept of competence has received increased attention in educational research in Germany. In particular, the “assessment of competencies plays a key role in optimizing educational processes and advancing educational systems” (Koeppen et al., 2008, p.

[8] Hasse, S., Joachim, C., Bögeholz, S. & Hammann, M. (2014). Assessing teaching and assessment competences of biology teacher trainees: Lessons from item development. *International Journal of Education in Mathematics, Science and Technology, 2*(3), 191-205.

61). Also, theoretical competence models (e.g., Bybee 1997) are presently being given an empirical foundation. Though current efforts in competence modelling and assessment have focussed on student competences mainly, teacher competences have also been closely studied. Teacher competences have received even more attention after the German Federal Ministry of Education and Research launched a funding initiative dedicated to the modeling and assessment of competences in higher education in 2012 (KoKoHS; cf. Blömeke & Zlatkin-Troitschanskaia 2013).

The present paper reports on a research project (ExMo) from this funding initiative. Its main focus is the development of measuring instruments geared at testing teaching competences and assessment competences of biology teacher trainees with regard to experimentation. Three German universities are involved in this project, i.e. University of Münster, University of Göttingen and University of Bamberg. As an intended effect, the measuring instruments are expected to contribute to improving science teacher education – an international request (European Commission 2011).

4.2 Theoretical Background and Rationale

4.2.1 Standards for Teacher Education in Germany

In the USA, there are several documents which focus on teaching standards in general and *Inquiry Teaching standards in detail* (NGSS 2013, NRC 1996/2000/2007, AAAS 1989). In Germany, comparable teaching standards are mostly lacking. While standards for teacher education exist, these standards are rather general and focus mainly on interdisciplinary and pedagogic competences. Specifically, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK 2004) has drafted a document with eleven standards for teacher education and training. These break down to aspects of Teaching, Education, Assessment and Innovation. An additional seven standards from this document pertain to biology lessons in particular. Merely one standard is devoted to Scientific Inquiry Teaching. In addition, the *Association for Subject Education* has published a framework for standards concerning the university phase of teacher training (GFD 2005). The document describes 20 standards in the following areas: Theoretical reflection of subject-matter education, subject-matter teaching, subject-specific assessment, subject-specific communication, development and evaluation of instruction and curricula. The standards also describe rather general aims such as: “Teacher trainees can describe and explain subject-specific educational concepts in a systematic way” (GFD 2005, p. 1).

Since teaching standards and assessment standards related to scientific inquiry are mostly lacking in Germany, it was necessary to specify the existing frameworks with respect to teaching scientific inquiry and assessing student achievement in scientific inquiry classes. Specifically, considerations were made concerning the question of what biology teacher trainees should be able to do (in terms of can-do statements) when they *analyze experimental biology lessons*, *plan experimental biology lessons* and *assess student achievement in experimental biology lessons*. Subsequently, test items related to these three dimensions were developed in order to build reliable and valid measures.

4.2.2 Teaching Experimentation in Biology Lessons

Internationally, science educators agree that scientific inquiry is central for the acquisition of scientific literacy. In addition, educational research has documented the contribution of experimental classroom experiences for the development of the leaners' scientific literacy (Abell 2007, Hofstein & Lunetta 2004, Sandoval & Reiser 2004, Chinn & Malhorta 2002, Psillos & Niedderer 2002).

Many countries have implemented teaching standards for scientific inquiry, which underlines the importance of scientific inquiry in general and of experimentation in particular (NGSS 2013, NRC 1996, AAAS 1993, Council of Ministers of Education 1996 [Canada], Department of Education 1995 [England], Ministry of Education 1993 [New Zealand], KMK 2004 [Germany]). However, learners are often unable to meet the expectations formulated in the standards (Grigg et al., 2007, Coble & Allen 2005, Bybee & Fuchs 2006, PISA 2004). Against this background, the National Research Council has argued that the learning outcomes need to be seen in the context of classroom teaching: "What students learn is greatly influenced by how they are taught" (1996, p.28).

Central ideas for effective scientific inquiry teaching are made explicit in the National Science Education Standards (NGSS 2013, NRC 1996). In Germany, the comparable documents are less detailed – as described above – and, as a consequence, they provide less guidance for teachers who intend to teach scientific inquiry in the classroom. However, scientific inquiry teaching in German schools often draws on the principles of inquiry teaching approaches that have been published internationally (cf. Hammann et al., 2008, Sandoval & Reiser 2004, Mulhall & Loughran 2003, Colburn 1997, White & Gunstone 1992). The following two examples are intended to illustrate this point.

In Germany, the national biology education standards (KMK 2004) specify that learners are expected to be able to *form hypotheses*, *plan experiments* and *analyze data*. These competences are theoretically grounded in the SDDS-Model (Scientific Discovery as

Dual Search) by David Klahr (2000). Biology teachers need to be able to support students in acquiring these competences, for example by following the recommendation that instruction mirror the phases that can be observed when scientists engage in scientific inquiry. Anderson states: “It is implied that inquiry learning should reflect the nature of scientific inquiry” (2002, p. 2). This recommendation can also be found in an important document issued at the beginning of a large national project for increasing the quality of science and mathematics education in Germany (Bund-Länder Kommission 1997).

Further, scientific inquiry can be used to teach *contents* and *methods*. The dual function of scientific inquiry is clearly visible in current approaches to teaching scientific inquiry, for example when learners are expected to “develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world” (Anderson 2002, p. 2). When students engage in experiments on seed germination, for example, they can learn about the factors responsible for this phenomenon, but also about the control-of-variable-strategy. Scientific inquiry teaching is thus marked by instructional measures that aim at a conceptual understanding as well as an understanding of the aims and methods of scientific inquiry.

Future biology teachers should be trained to take these exemplary ideas and distinctions into consideration when planning and analyzing experimental biology lessons. These ideas and distinctions are also central for developing a measurement instrument that aims at testing teacher trainees’ competences, as the two following examples show:

- In a test item concerned with assessing the competence of planning experimental biology lessons, a work sheet is depicted that a teacher wants to use in class. In the work sheet, the phase of hypothesis formation is not taken into account. Thus, the work sheet it is not systematically oriented towards the stages of scientific inquiry. The teacher trainees are asked to modify the work sheet in a way that it also promotes hypothesis formation.
- In a test item concerned with assessing the competence of analyzing experimental biology lessons, a situation is depicted where a group of learners records data that contradicts scientific findings. The teacher considers excluding the data of this group based on the rationale that incorrect data does not promote an adequate understanding of a biological phenomenon. The teacher trainees are asked to decide whether or not the teacher’s intended action is appropriate.

> The teacher trainees are expected to recognize that it is not content knowledge alone that can be gained from an experiment. Disconfirming data can also be used to train students how to analyze data appropriately.

Item development very soon made it clear that there are multiple alternative ways to proceed when doing scientific inquiry and that it is impossible to expect teacher trainees to describe *the one and only* correct way. Item development, as indicated above, built on the idea that there are more or less effective ways of teaching scientific inquiry – and that mismatches between educational goals and procedures must be avoided, but this does not mean "that all teachers should pursue a single approach to teaching science" (Anderson 2002, p.2).

4.2.3 Definition of Competences

In this paper, the focus lies on teachers' *competences*, e.g., analyzing experimental biology lessons, planning experimental biology lessons and assessing student achievement in experimental biology lessons. Drawing on Weinert (2001), Klieme & Leutner (2006) and Koeppen et al. (2008), competences are defined as "context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains" (Koeppen et al., 2008, 62).

Specifically, the competence to analyse lessons is defined as the cognitive disposition to "appropriately apprehend and assess the quality of observed lessons with regard to effectiveness" (Plöger & Scholl 2014).

Further, the competence to plan lessons is defined as the cognitive disposition to "anticipate goal-oriented actions in future situations. It is connected to the determination of prerequisites for successful actions (e.g., learning preconditions of students or the availability of materials, media, tasks) and to the thinking through of different opportunities for action in order to decide on a certain course of action" (Kiper 2012).

Finally, the *competence of assessing student achievement* is considered as the cognitive disposition to "continuously assess the level of knowledge, learning progress and performance difficulties of individual learners as well as the difficulties of different learning tasks" (Weinert 2000, p.14).

4.3 Target Group

The study described in this paper aims at assessing the competences of university students intending to become biology teachers. Future biology teachers decide at the beginning of their university studies, which teaching certificate they aim for: (i.e., high school, comprehensive school, vocational school and academic high school.) All types of biology teachers were included. Also, the sample included students from the two phases of university education (BA and MA). Several German universities from the Länder of North Rhine-Westphalia, Lower Saxony, Mecklenburg-Hither Pomerania and Bavaria participated in the pre-piloting and piloting of the measurement instruments.

4.4 Considerations for the Development of the Measurement Instruments

4.4.1 Connection to current research

Pedagogical Content Knowledge and Competence: In the USA and in many countries world-wide, teachers' expertise is currently being researched within the framework of Pedagogical Context Knowledge (PCK). A European contribution to PCK research is its emphasis on teachers' competences –rather than teachers' knowledge – a difference that will be further elaborated in the following part of the paper.

American researchers assume a *knowledge base of teaching* (Shulman 1986, 1987), which consists of several *categories of knowledge*, including Pedagogical Content Knowledge. The dimensions of PCK are framed differently depending on the research group. Shulman (1986, 1987), for example, names seven categories of PCK relevant for science teaching, Magnusson et al. (1999) five. The term *knowledge* seems to be the focal point of American research.

German research regarding teachers' professional knowledge utilizes the framework of international PCK research, but focusses on assessing competence. The terms *knowledge* and *competence* refer to different constructs. The term competence is defined as the "mental conditions necessary for cognitive, social and vocational achievement" (Weinert 1999, p. 26). Thus, the emphasis lies on coping with real-world problems. As a consequence, competence research focuses on problem solving skills, i.e., "all those skills required to evaluate the relevant features of a problem, so that suitable solution strategies can be selected and used" (Weinert 1999, p. 8). Without PCK however an instructor cannot be competent. "Knowledge is the necessary foundation of competence" (Weinert 1999, p.5).

PCK-models, hence, are not identical with competence models. Rather, competence models focus on a defined psychological construct (see above) and they specify the structures of a competence (structure models), levels of competence (stage models) and changes in competence through instruction and in time (development models) (cf. Koeppen et al., 2008). Structural similarities, however, can be seen, when the components / categories of PCK models are compared to the structure model of teacher trainee competences presented in this paper (i.e., analyzing experimental lessons, planning experimental lessons and assessing student achievement in experimental lessons). Specifically, it is possible to draw on the PCK-model by Magnusson et al. (1999) in order to illustrate similarities. In Magnusson's model, five components of PCK are described: *Orientation to Teaching Science, Knowledge of Science Curricula, Knowledge of Assessment of Scientific Literacy, Knowledge of Instructional Strategies* and *Knowledge of Students´ Understanding of Science*. The competences of analyzing and planning experimental lessons can be attributed to the PCK-components of *Knowledge of Students´ Understanding of Science* and *Knowledge of Instructional Strategies*. Further, the competence of assessing student achievement in experimental lessons can be related to the PCK component of *Knowledge of Assessment of Scientific Literacy*.

Projects with related Objectives: Test instruments for assessing the competences of planning and analyzing lessons focusing on scientific inquiry are rare. Prior to this project, however, it was possible to find related studies with similar research questions.

The project *Pedagogy of Science Inquiry Teaching Test* (POSITT, Cobern et al., 2014) is concerned with assessing pedagogical content knowledge of inquiry science teaching. The POSIT-Test is an important reference point for the present study, as item development for POSITT showed that it is possible to use realistic vignettes with questions related to them for a paper-and-pencil test. A similar approach to item development is presented in this paper. POSITT, however, focuses on teacher trainees' preferences regarding different teaching strategies and, assesses so-called *teachers´ orientations*. In ExMo, in contrast, realistic teaching vignettes are used in order to assess teachers´ competences.

In the project *Professional Minds,* Oser (2010) examines the quality of complex competence profiles (not individual competences) of teachers, which include cognitive aspects (e.g. *clarity of task*) as well as affective aspects (e.g. *acceptance*, *empathy*). ExMo, in contrast focusses on individual competences which are defined as cognitive dispositions.

Teachers' analyzing competence is currently being investigated in a project by Plöger and Scholl (2014). This study, however, is not concerned with a specific, subject-specific procedural competence like experimentation. Instead more universal aspects related to analyzing classroom situations are being examined. Plöger & Scholl (2014) use the model of *hierarchical complexity* (Commons 2008), and distinguish between *horizontal complexity* and *vertical complexity. The same framework* is also used in the study presented here for developing items and for coding the answers (see *options for the coding of open tasks*, p.7).

Seidel et al. (2011) investigate teachers' perception of classroom situations. Specifically, classroom situations are presented in the form of video vignettes and teachers are asked to analyze them. In this study, rather general criteria (as opposed to subject-matter specific criteria) are used, such as e.g. the difference between describing and explaining a classroom situation. The distinction, however, is well taken. It is relevant for item development in the study presented here. The concept of professional perception (Goodwin, 1994; Sherin, 2002) states that the mere description of a lesson puts lower requirements on a teacher than explaining and predicting. This aspect of analysing a lesson is taken into account in ExMo for the development of tasks and code manuals as well.

Baer et al. (2011) investigate teacher trainees' knowledge about important aspects of planning a lesson. The focus of their research is the teacher trainee's knowledge of important concepts (de Jong & Ferguson-Hessler, 1996), for example knowledge of teaching methods and curricula. The level of specificity, however, required for answering the items, is very general. The teacher trainees, for example, can solve an item by simply stating that it is important to plan longer teaching units (as opposed to individual lessons) and that it is important to make choices against the background of their knowledge of curricula. Also, subject-matter specific aspects regarding experimentation are not taken into account in this project.

Dübbelde (2013) examines diagnostic competences of biology teacher trainees concerning the domain of knowledge acquisition. The project aims at developing a test instrument with closed task types for status and process diagnostic competences. Among other things it is recorded how far biology teacher trainees assess students' results and work processes when experimenting with the help of given evaluation criteria. Dübbelde pursues a partly similar aim as we do within ExMo regarding assessment competences. In her test instrument teacher trainees are given, for instance, a worksheet filled out by two students to document the steps of their experiment. The teacher trainees are asked

to assess the students' results with regard to the given criteria. For each criterion, the teacher trainees have to choose one of three (or four) alternative answers. For instance, they have to assess whether the students' hypothesis is related to the research question by ticking off "*yes"*, "*no"* or "*don't know"*.

The test instrument used in Dübbelde (2013) includes comparable criteria pertaining to experimentation as the ExMo test instrument. In ExMo, however, it is of central interest to find out to what extent the teacher trainees know (and activate on their own) criteria with respect to experimentation, typical preconceptions and difficulties students have when experimenting. In addition, we are interested in knowing to what extent teacher students are able to independently utilize these for the assessment of students' achievements. For a differentiated evaluation of the teacher trainees' assessment cognitions open tasks are used in ExMo. The tasks describe students´ performance in experimenting and then ask the teacher trainees to assess either the formation of hypotheses, planning of experiments or data analysis. For this, the teacher trainees have to be aware of the criteria and apply them correctly and in a sophisticated manner.

4.5 Selection of Subject-Specific Content

The teaching vignettes focus on biological topics that can be found in the curricula of most Länder in Germany. Also the biological topics chosen can be combined with experiments pertinent to students. For the grades 5-6, seed germination was chosen, for grades 7-8 photosynthesis and for grades 9-10 enzymes.

4.6 Central Challenges

First attempts at item development quickly showed two major challenges, which deserve closer study:

1) In order to assess the teacher trainees' competence to analyze lessons, the complexity of the situation has to be reduced to some degree so that it is possible to code the answers of the teacher trainees´ test. At the same time, the complexity shouldn't be reduced too far so that the realistic character of classroom situation doesn't get lost. The aim is to assess a person's competence to solve real-world problems *and* arrive at answers that can be coded.

2) In order to assess the competence of planning lessons, the openness of planning decisions must be restricted to some degree in order to arrive at answers that can be coded. However, the situation should not be reduced too much,

so that the character of the situation still classifies as real-world problem solving. This situation is analogous to the situation described under challenge 1.

4.6.1 Dealing with Challenge #1

In order to sufficiently reduce the complexity of analyzing classroom situations, the decision was made to explicitly state which competence the teacher in the teaching vignette intends to promote when teaching an experimental lesson. Also, the question that needed to be answered was framed in a way that decreased the possibility of variation. In a current item (see Appendix: *Task 1*), the description can be found that a teacher has three different options in order to promote the student competence of planning experiments independently. The teacher trainee's task is to judge which approach is the most suitable and give reasons for their decision.

During item development two further insights were gained: Multiple choice questions proved unsuitable because it was found possible to answer them through logical reasoning and reading skills alone (see Appendix: *Task 2*). Also, open-answer tasks, which did not specify the competence the teacher intends to promote, allowed for too much variation in answers so that coding the answers proved impossible.

4.6.2 Dealing with Challenge 2

Similar to challenge 1, it was necessary to find a way of limiting the variation in possible answers. In particular, the item contains a description of an experimental lesson. The teacher trainees are encouraged to plan alternatives or suggest changes because specific aspects of the plan contain flaws or mismatches between intended aims and specific aspects of the lesson. Generally, items assessing the competence to plan experimental lessons, also state which experimental competence the teacher intends to promote.

4.7 Options of Coding Open Tasks

When coding the answers we utilized Commons' (2008) concept of complexity. Commons describes that it is possible to distinguish complexity in two ways: *Horizontal complexity* implies that several pieces of information are processed on the same level, while *vertical complexity* entails a processing of information on different levels. With regard to teaching and assessing competences of teachers, this model of complexity can be applied as follows: When analyzing, planning and assessing, teachers must constantly take several unrelated aspects into account. This may entail e.g. aspects related to subject matter, social aspects and methodological teaching aspects. A teacher has to consider several students´ conceptions that are independent from each other or diagnose student

errors, which occur simultaneously but independent from each other (=horizontal complexity). The more aspects there are that need to be considered, the greater is the challenge for the teacher. It is not only the amount of tasks to be managed simultaneously but also the difficulty of an individual task, which influences the complexity of the challenge. Thus it is easier e.g. to simply name an occurring problem rather than give a well-founded explanation of the causes of the problem (=vertical complexity).

An exemplification of the coding manual of a task that encompasses both horizontal and vertical complexity can be found in the Appendix (see *Task 1).*

At the end of the task, teacher trainees are required to rank three options from the easiest to the most difficult and to describe which aspects of planning an experiment are responsible for the different levels of difficulty. The following three aspects can be distinguished for differentiating between the difficulty of the three options (cf. Hammann et al., 2007):

1. Does the teacher tell the students which factors need to be examined [easier] or do the students have to determine the factors themselves [harder]?
2. Do the students have to examine one factor [easier] or a several factors [harder]?
3. Do the students have to plan a small number [easier] or a large number [harder] of experimental setups?

The maximum score for this task is 4 points. Mentioning the three difficulty-generating aspects (horizontal complexity) and giving reasons for the three difficulty-generating aspects (vertical complexity) are scored with one point each, as is the correct ranking of the three aspects. The assumption underlying this coding is that on the one hand a teacher needs well-founded theoretical knowledge about the difficulty-generating aspects of experiment planning while on the other hand especially the performance during the lesson is key for students' learning success. Hence, a teacher trainee who names the correct and consequently sensible order for the practical application during a lesson, but only names two of the difficulty-generating aspects receives the same number of points as a teacher trainee who names all three aspects but does not arrange the options in an appropriate way.

The coding guide provides guidelines as well as anchor examples and contrasting examples, as specified by Bühner (2011).

4.8 Item Development

4.8.1 Iterative Process

According to Wilson (2005), item development is an cyclical process with four "*building blocks"* (i.e., *construct maps, item design, outcome space and measurement model*). The results of each step in the process inform the next step. Also, the process is iterative and the cycle may be repeated multiple times.

4.8.2 Item Development for nine Facets of Teaching and Assessment Competence

Nine facets (see Table 4-1) arise as a result of crossing three teachers' competences with three student competences. Prior to item development, a framework for item development was drafted in order to provide a systematic basis that was meant to ensure the subsequent comparability of all tasks in data analysis (Murphy & Davidshofer, 2005; Gruijter 2008). This framework states, for example, that the item development follows the approach of rational item construction (Kline, 2005), that items require open-responses, and that items start with a description of a realistic situation.

Table 4-1 Facets of teaching experimentation in biology

Teachers competences / *Students competences*	*Analyzing experimental lessons* Analyzing teachers´ decisions that aim at ...	*Planning experimental lessons* Planning instructions that aim at ...	*Assessing students achievements in experimental lessons* Assessing the quality of ...
Forming hypotheses	...teaching students how to form hypotheses	...teaching students how to form hypotheses	...hypotheses formed by students
Planning experiments	...teaching students how to plan experiments	...teaching students how to plan experiments	...experiments planned by students
Analyzing data	...teaching students how to analyze data	...teaching students how to analyze data	...students interpretations gained by analyzing data

4.9 Formulation of concrete Requirements for Teaching Experimentation

In general, the development of a test for assessing complex features must always be preceded by a specification of the object of measurement (cf. Kline, 2005). Taking into consideration the relevant specialized literature (e.g., Carey et al., 1989; White & Gunstone 1992; Gott & Duggan, 1995; Driver el al., 1996; Colburn, 1997, Chen & Klahr, 1999, Kanari & Millar, 2004, Bybee et al., 2006, Hammann et al., 2008; Ford, 2008;

Gyllenpalm et al., 2010), central requirements for biology teachers when teaching experimentation were organized with regard to the nine facets.

The latter shall be illustrated by means of an example for the facet of *Analyzing teachers´ decisions that aim at teaching students how to form hypotheses*: A biology instructor should be able to...

- …evaluate and analyze the challenges in planning different experimental courses of action.
- …identify the aspects that constitute the range of complexity of different tasks. This especially includes the number of variables to be tested and the number of experimental setups to be compared as well as naming of the variables to be tested.

This concrete requirement was operationalized in the test item discussed above (see Appendix: *Task 3*).

4.10 Discussing Prototypical Tasks with Experts

Following the development of prototypical items, a multi-day workshop was conducted. During this workshop a framework for the item development and prototypical items were introduced and discussed. As part of this meeting all prototypical tasks were discussed, modified or excluded, if they proved unsuitable for the assessment of the targeted competence.

In addition, the tasks and items for the evaluation of assessment competence were tested in an expert panel for the validity of their content. Six experts (among them three scientists and three teachers) came to the conclusion that the lesson vignettes can be considered realistic and the tasks may be considered part of the interesting collectivity of possible tasks for assessment competence regarding experimentation. The results of this expert survey were taken into account in the further development of items.

4.11 Studies of Thinking-Aloud Protocols

The aim of think aloud protocols (cf. Ericsson & Simon, 1980 & 1999) is to assess people's' cognitive processes (Hussy et al., 2010), for example in order to make sure that the items are suited to initiate the processes that are expected to occur when analyzing a lesson, planning a lesson and assessing learning outcomes.

In the study, 32 biology teacher students (16 people worked on items concerned with analyzing and planning at the University of Münster, an additional 16 people worked on

items concerned with assessment at the University of Göttingen) we presented with 8 or 10 items each. All test persons took part in the study individually and received a standardized methodological instruction to the study of thinking aloud in the beginning. The think aloud protocols were analyzed qualitatively in order to refine items for the following quantitative studies.

4.12 Item Piloting and Analysis

4.12.1 Sample and Goals

The piloting of the developed tasks encompassed 2 subsequent studies: In the pre-pilot, 51 students of the Universities of Münster and Göttingen participated. In total 60 items were tested. Each teacher trainee received a test booklet with 9 items that either required analyzing and planning experimental lessons (N=27) or assessing student achievement in experimental lessons (N=24). The aims of the pre-pilot were the advancement of the scoring guides and the optimization of tasks.

In the second study, the pilot study, 160 students from six German universities have participated so far. In this phase, each testing booklet contains 9 items concerning analysis and planning or the assessment of students' achievements.

4.13 Work so far

The project ExMo currently moved on to its second pilot stage. The completion of assessment and a thorough analysis of the data, which allows for analyses of reliability and validity of the testing instrument, are still pending. When the data is available, a comparison between Bachelor and Master's students will be conducted in order to investigate whether Bachelor students have less developed competences than Master students. Should this be the case it will be considered indicative of acquirable cognitive competences having been measured rather than intelligence. The preliminary results of the study with thinking aloud indicate that the competences increase over the courses of university education and that students acquiring a teaching degree for academic high school perform better than students acquiring a teaching degree for any other school type. These findings are descriptive and explorative and they were not statistically tested.

4.14 Conclusion

Requirements for the development of paper-and-pencil tasks were described with respect to the assessment of teaching competences (analysis and planning of lessons). Specifically, lessons from item development showed that it is necessary to restrict the

openness of the planning situation to a degree where it is possible to code whether the planning decision was made on the basis of subject-matter specific knowledge. Furthermore, it is necessary to specify the learning objectives when assessing analyzing competence so far to allow judgment on whether or not the given classroom scenarios were appropriately analyzed.

The approach seems promising despite it being impossible to report on inter-rater agreement, reliability and validity at this point. It is presumably possible to transfer the principles of item development to other subject-educational contexts, e.g. the planning and analysis of non-experimental biology lessons. Science educators, who are interested in the measuring of competences, are encouraged to test this approach and apply it to other areas.

References

AAAS. (1993). *Benchmarks for science literacy*. Washington D.C.: American Association for the Advancement of Science.

Abell, S. K. (2007). Research on science teacher knowledge. In S. K. Abell, & N. G. Lederman (Eds.), *Research on science teacher education* (pp. 1105-1149). New York: Routledge.

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1-12.

Baer, M., Kocher, M., Wyss, C., Guldimann, T., Larcher, S., & Dörr, G. (2011). Lehrerbildung und Praxiserfahrung im ersten Berufsjahr und ihre Wirkung auf die Unterrichtskompetenzen von Studierenden und jungen Lehrpersonen im Berufseinstieg. *Zeitschrift für Erziehungswissenschaft, 14*(1), 85-117.

Blömeke, S. & Zlatkin-Troitschanskaia, O. (Eds.) (2013). *The German funding initiative "Modeling and Measuring Competencies in Higher Education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students.* (KoKoHs Working Papers, 3). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung. (1997). Gutachten zur Vorbereitung des Programms "Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts". Bonn.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte Auflage). München: Pearson Studium.

Bybee, R. W., Taylor, J. A., Gardner, A., van Scotter, P., Carlson Powell, J., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins and effectiveness.* Unpublished manuscript.

Bybee, R. W., & Fuchs, B. (2006). Preparing the 21st century workforce: A new reform in science and technology education. *Journal of Research in Science Teaching, 43*(4), 349-352.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). "An experiment is when you try it and see if it works": A study of grade 7 students´ understanding of the construction of scientific knowledge. *International Journal of Science Education, 11*(special issue), 514-529.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*(2), 175-218.

Coble, C. R., & Allen, M. (2005). *Keeping America competitive: Five strategies to improve mathematics and science education*. Denver: Education Commission of the States.

Cobern, W. W., Schuster, D. G., Adams, B., Skjold, B., Mugaloglu, E. Z., Bentz, A., & Sparks, K. (2014). Pedagogy of Science Teaching Tests: Formative Assessments of Science Teaching Orientations. *International Journal of Science Education*. http://bit.ly/RE95xZ

Colburn, A. (1997). How to make lab activities more open ended. *CSTA Journal,* (Fall 1997), 4-6.

Commons, M. L. (2008). Introduction to the model of hierarchical complexity and its relationship to postformal action. *World Futures, 64*(5-7), 305-320.

Council of Ministers of Education. (1996). *Common framework of science learning outcomes K-12 (draft)*. Victoria B.C.: Ministry of Education, Skills and Training.

de Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*(2), 105-113.

Department of Education. (1995). *Science in the national curriculum*. London.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Buckingham u.a.: Open Univ. Press.

Dübbelde, G. (2013). *Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung*. http://nbn-resolving.de/urn:nbn:de:hebis:34-2013122044701.

Ericsson, K. A., & Simon, H. A. (1999). *Protocol analysis: Verbal reports as data* (Rev, 3 print ed.).Cambridge, Mass. u.a.: MIT Press.

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

European Commission (2011). In Eurydice (Ed.), *Naturwissenschaftlicher Unterricht in Europa*. Brüssel: Exekutivagentur Bildung, Audiovisuelles und Kultur.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 93*(3), 404-423.

Gesellschaft für Fachdidaktik e.V. (2005). *Fachdidaktische Kompetenzbereiche, Kompetenzen und Standards für die 1. Phase der Lehrerbildung (BA+MA)*

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*(3), 606-633.

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

Grigg, W., Donahue, P., & Dion, G. (2007). In US Department of Education (Ed.), *The nation's report card:12th-grade reading and mathematics*, 2005. NCES 2007-468. Washington D.C.: ERIC.

Gruijter, Dato N. M. de, & Kamp, L. J. T. v. d. (2008). *Statistical test theory for the behavioral sciences*. Boca Raton u.a.]: Chapman & Hall/CRC, Boca Raton u.a.].

Gyllenpalm, J., Wickman, P., & Holmgren, S. (2010). Teachers´ language on scientific inquiry: Methods of teaching or methods of inquiry. *International Journal of Science Education, 32*(9), 1151-1172.

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

Hammann, M., Ganser, M., & Haupt, M. (2007). *Experimentieren können. kompetenzentwicklungsmodelle und ihre nutzung im unterricht*. Geographie Heute, (255/256), 88-91.

Hammann, M.,Phan, T. T. H., Ehmer, M., & Bayrhuber,H. (2006). Fehlerfrei experimentieren. *Mathematischer und Naturwissenschaftlicher Unterricht, 59*(5), 292-299.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28-54.

Hussy,W., Schreier, M., & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor*. Berlin: Springer.

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.

Kiper, H. (2012). *Unterricht planen, durchführen, auswerten - Überlegungen zur lernwirksamen Unterrichtsplanung*. In K. Bauer, & N. Logemann (Eds.), (pp. 151-182). Münster: Waxmann Verlag.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge: MIT Press.

Klieme, E., & Leutner, D. (2006). *Kompetenzmodelle zur erfassung individueller lernergebnisse und zur bilanzierung von bildungsprozessen. beschreibung eines neu eingrichteten schwerpunktprogramms der DFG* [competence models for assessing individual learning outcomes and evaluating educational processes. description of a new priority program of the German research foundation, DFG]. Zeitschrift Für Pädagogik, 52(6), 876-903.

Kline, T. (2005). *Psychological testing*. Thousand Oaks: SAGE.

KMK. (2004). Bildungsstandards im Fach Biologie für den mittleren Schulabschluss. Beschluss vom 16.12.2004 Luchterhand.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. Zeitschrift Für Psychologie/*Journal of Psychology, 216*(2), 61-73.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome, & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95-132). Dordrecht: Kluwer Academic Publishers.

Ministry of Education. (1993). *Science in the New Zealand curriculum*. Wellington, New Zealand: New Media.

Mulhall, P., Berry, A., & Loughran, J. (2003). Framework for representing science teachers´ pedagogical content knowledge. *Asia-Pacific Forum on Science Learning and Teaching, 4*(2)

Murphy,K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6, international ed.). Upper Saddle River, NJ: Pearson Education International, Prentice-Hall.

National Research Council. (2000). *Inquiry and the national science education standards. A guide for teaching and learning*. Washington D.C.: National Academy Press.

National Research Council. (2007). In Duschl R. A., Schweingruber H. A. and Shouse A. W.(Eds.), *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academies Press.

National Research Council. (1996). *National science education standards*. Washington D.C.: National Academy Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington D.C.: The National Academy Press.

Oser, F., Heinzer, S., & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. Chancen und Grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft, 38*(1), 5-28.

PISA-Konsortium Deutschland. (2004). PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland– Ergebnisse des zweiten internationalen Vergleichs.

Plöger, W., & Scholl, D. (2014). Analysekompetenz von Lehrpersonen – Modellierung und Messung. *Zeitschrift für Erziehungswissenschaft, 17*(1), 85-112.

Psillos, D., & Niedderer, H. (2002). *Teaching and learning in the science laboratory*. Dordrecht: Kluwer Academic Publishers.

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

Sherin, M. G. (2002). When teaching becomes learning. *Cognition and Instruction, 20*(2), 119-150.

Shulman, L. S. (1986). Those who understand: Knowledge growths in teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Seidel, T, Stürmer, K, Blomberg, G, Kobarg, M, Schwindt, K. (2011). “Teacher learning from analysis of videotaped classroom situations: does it make a difference whether teachers observe their own teaching or that or others?” *Teaching and Teacher Education, 27*: 259-267.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen, & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45-65). Seattle: Hogrefe & Huber Publishers.

Weinert, F. E. (2000). Lehren und Lernen für die Zukunft – Ansprüche an das Lernen in der Schule. *Pädagogische Nachrichten Rheinland-Pfalz 2*, 1-16

Weinert, F. E. (1999). Konzepte der Kompetenz. Gutachten zum OECD-Projekt "Definition and Selection of Competencies: Theoretical and conceptual foundations (DeSeCo)".

White, R., & Gunstone, R. (1992). *Probing understanding*. London: Falmer Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. London: Routledge.

Appendix

Task 1: Item to assess the competence Analyzing experimental lessons related to the teaching objective planning experiments (current version)

Analysing experimental lessons — Planning experiments

Mr. Hahn teaches biology in a sixth grade. He started working on seed germination and would like his students to experiment actively in class. Mr. Hahn wants to focus on the promotion of the competence to **plan experiments**.

As the learning group does not have many experiences in experimenting independantly, he compares three different possible approaches to teach students how to plan experiments (see A-C).

He considers what the students are required to do in each of the three options. He would like to rank them from the easiest to the hardest concerning the difficulty of planning experiments.

Three approaches to initiate the planning of experiments

A) The teacher hands out bean seeds to the students. The learners are instructed to find out by which factors the seed germination is affected.

B) The teacher shows the students a flower pot with soil containing beans that has germinated and describes the precise conditions of the germination. He hands out bean seeds to the students. The students are instructed o find out whether or not the soil is required for seed germination.

C) The teacher hands out bean seeds to the students. The students are instructed to find out whether or not seeds require light and warmth for germination. In addition the students are instructed how to handle with factors other than light and warmth.

Task:

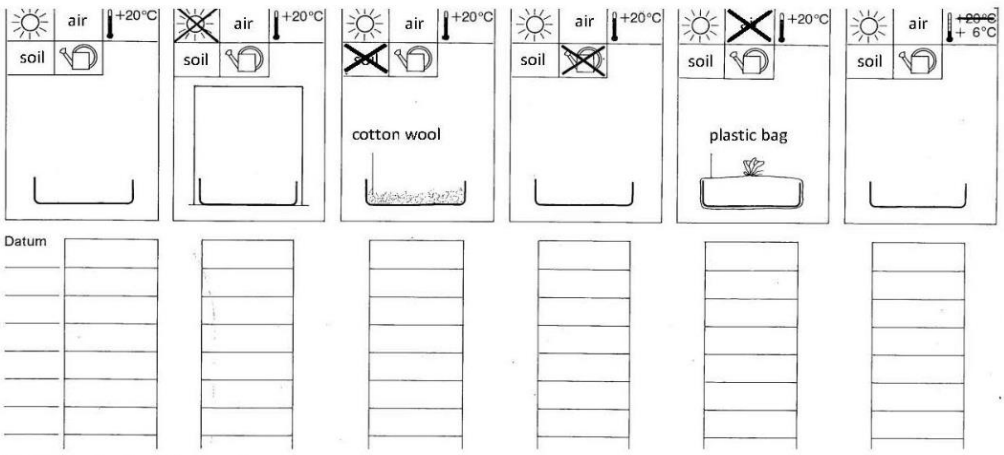
Rank the three options from the easiest to the hardest concerning the difficulty of planning experiments.

Analyze which aspects determine the level of difficulty of the three approaches to teaching students how to plan experiments!

Task 2: Multiple choice – Item to assess the competence *analyzing experimental lessons*

A teacher wants to improve the students´ competencies of carrying out biological experiments by using the worksheet depicted below.

Essential factors for the germination of bean seeds:
1. Put 10 bean seeds into each petri dish and keep it under the stated conditions
2. Make observations every second day about the germination ("+" for each germinated seed "–" for each not germinated seed)
3. Write down the conclusions of your experiment and complete the answer sentences



Datum

For the germination of bean seeds

These factors are essential: _____

These factors are NOT essential:

Task:

Which of the following competencies will be fostered by answering the three tasks on the worksheet?

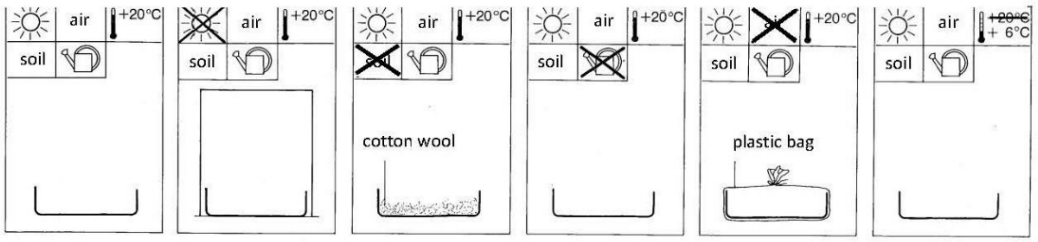
	Check the correct answers
Asking scientific questions	( )
Forming hypotheses	( )
Planning experiments	( )
Keeping precise records	( )
Analyzing data	( )
Analyzing experiments critically	( )
Relating results to new questions	( )

Task 3: Open format item to assess the competence *analyzing experimental lessons*

A teacher wants to improve the students´ competencies of carrying out biological experiments by using the worksheet depicted below.

Essential factors for the germination of bean seeds:

1. Put 10 bean seeds into each petri dish and keep it under the stated conditions
2. Make observations every second day about the germination ("+" for each germinated seed " –" for each not germinated seed)
3. Write down the conclusions of your experiment and complete the answer sentences



Datum

For the germination of bean seeds

These factors are essential: _____

These factors are NOT essential:

Task:

Are the three tasks on the worksheet suitable to improve the pupil's competencies of conducting experiments? Give reasons for your assertions!

5. Kompetenzen von (angehenden) Biologielehrkräften zur Beurteilung von Experimentierkompetenzen[9]

Competences of Biology Teacher Students to Assess Experimental Competences

Eine zentrale Erkenntnismethode im Biologieunterricht ist das Experimentieren. Für den Experimentalunterricht sind entsprechende Lehrerkompetenzen gefordert. Der ExMo-Verbund befasst sich mit der empirischen Fundierung eines Kompetenzmodells zu Vermittlungs- und Beurteilungskompetenzen zum Experimentieren für (angehende) Biologielehrkräfte. Berichtet wird über einen methodischen Ansatz zur Aufgabenentwicklung und -auswertung zur postulierten Teilkompetenz Beurteilungskompetenz für Schülerexperimentierkompetenzen. In einer Pilotstudie (N = 145 Biologielehramtsstudierende von acht bundesdeutschen Universitäten) wurden systematisch entwickelte Aufgaben in verschiedenen Testheften (incomplete block design) erprobt. Die Auswertungen erfolgten mit Blick auf Reliabilität und Validität. Die Ergebnisse geben – bei allen Grenzen der Studie – empirische Hinweise zur Tragfähigkeit des gewählten Ansatzes zur Kompetenzmessung. Abschließend wird ein Ausblick auf weitere Validierungsvorhaben im Zusammenhang mit dem Kompetenzmodell gegeben.

Schlüsselwörter: Kompetenzmessung, Experimentieren, Lehrerkompetenzen, Kompetenzmodell

Experimentation is a key method of scientific inquiry. Future biology teachers need to acquire specific teacher competences to be prepared for this aspect. In the joint research project ExMo, we aim at developing an empirically tested competence model regarding “teaching competences and assessment competence in experimental biology lessons” for biology teacher students. First, we describe the methodological approach to item development and to coding of responses by focusing on assessment competence. We report on a pilot study (N = 145 biology teacher students from eight German universities) testing a systematically developed range of items organized in test booklets (incomplete block design). Then, we introduce the findings regarding reliability and validity. Our findings indicate that the approach of assessment of competences is promising given the

[9] Bögeholz*, S., Joachim*, C., Hasse, S. & Hammann, M. (2016). Kompetenzen von (angehenden) Biologielehrkräften zur Beurteilung von Experimentierkompetenzen. *Unterrichtswissenschaft, 44*(1), 40-54. (*gemeinsame Erstautorenschaft)
This is a post-peer-review, pre-copyedit version of an article published in Unterrichtswissenschaft.

limitations of the study. Finally, we present an outlook on the further validation of the competence model.

Keywords: assessment of competences, experimentation, teacher competences, competence model

5.1 Einleitung

Seit über zehn Jahren wird dem *Beurteilen* ein eigenständiger Kompetenzbereich in den Standards für die Lehrerbildung gewidmet (*Kultusministerkonferenz* [KMK], 2004a). Spezifische Beschreibungen für Anforderungen folgten mit dem Kompetenzbereich *Fachbezogenes Diagnostizieren und Beurteilen* durch die Gesellschaft für Fachdidaktik. Der Kompetenzbereich umfasst u. a. die Fähigkeit, „Modelle und Kriterien […] der Beurteilung auf fachliches Lernen zu beziehen“ (*Gesellschaft für Fachdidaktik* [GFD], 2005a). Dazu zählt z. B. die „Kenntnis von Kompetenzmodellen und Standarddefinitionen sowie […] Methoden zur Erfassung und Beurteilung von Schülerleistungen“ (B3 in GFD, 2005b).

Während Schülerkompetenzen seit spätestens der Jahrtausendwende stark im bildungswissenschaftlichen Fokus stehen (z. B. PISA), ist eine systematische Untersuchung von Lehrerkompetenzen vergleichsweise jung (z. B. im *Bundesministerium für Bildung und Forschung* [BMBF]-Verbund *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor* [KoKoHs]).

Betrachtet man den Unterricht in den naturwissenschaftlichen Fächern, so ist Erkenntnisgewinnung von großer Relevanz (KMK, 2004b). In diesem Bereich liegen aber auch spezielle Defizite vor: „Die Prinzipien selbst einfachster experimenteller Anordnungen verstehen bis zum Ende der 8. Klasse […] nur etwa 10-15 Prozent eines Jahrgangs“ (Baumert & Lehmann, 1997, S. 86). Berücksichtigt man, dass Diagnostik für systematische Förderung entscheidend ist (Hesse, 2014), kommt der Beurteilungskompetenz von Lehrkräften im Bereich Schülerexperimentierkompetenzen eine besondere Rolle zu.

5.2 Theoretischer Hintergrund

5.2.1 Bedeutung und Struktur von Schülerexperimentierkompetenzen

Die Bedeutung des Experimentierens wird in den normativen Vorgaben für den Biologieunterricht der Sekundarstufe I und II hervorgehoben. Eingefordert werden folgende Kom-

petenzen: Schülerinnen und Schüler „entwickeln Hypothesen, planen Experimente, führen diese durch und werten sie hypothesenbezogen aus“ (Nieders. Kultusministerium, 2009, S. 17; vgl. KMK, 2004b). Das Experimentieren erlaubt, Schlussfolgerungen über kausale Zusammenhänge – also über Beziehungen zwischen Ursache und Wirkung – zu ziehen. Dies erfolgt mittels eines „planmäßige[n], systematische[n] und zielgerichtete[n] sowie kontrollierte[n] Eingriff[s]“ (Meier & Wellnitz, 2013, S. 6). Damit stellt das Experimentieren eine zentrale Erkenntnismethode dar.

Hammann (2004) beschreibt – aufbauend auf dem *Scientific Discovery as Dual Search* Modell (Klahr, 2000) – Experimentierkompetenzen von Lernenden unterschiedlicher Klassenstufen und verweist auf Schwierigkeiten von Schülerinnen und Schülern beim Experimentieren. Die Teilkompetenzen *Suche im Hypothesen-Suchraum*, *Suche im Experimentier-Suchraum* und *Analyse von Daten* wurden in vier Niveaus graduiert. Lernende auf Niveau I experimentieren ohne Hypothesen, variieren die potenziell einflussreichen Faktoren unsystematisch und beziehen ihre Daten nicht auf eine vorab aufgestellte Hypothese. Demgegenüber suchen Schülerinnen und Schüler auf Niveau IV systematisch nach Hypothesen. Außerdem sind sie in der Lage, Hypothesen zu revidieren, den Einfluss von Variablen systematisch zu testen – auch in unbekannten Domänen – und Daten hypothesenbezogen auszuwerten (Hammann, 2004). Typische Schülerfehler beim Experimentieren sind z. B. das Aufstellen empirisch nicht testbarer Hypothesen, das Experimentieren im *Ingenieursmodus*[10] oder der *confirmation bias*[11] (z. B. Dübbelde, 2013; Hammann, 2004; Schmiemann & Mayer, 2013). Die hier beschriebenen Schülerkompetenzen und Schülerfehler beziehen sich auf kognitive Fähigkeiten, d. h.: *practical skills* werden in diesem Zusammenhang nicht fokussiert.

5.2.2 Lehrerkompetenzen: Beurteilungs- und Diagnosekompetenzen für Experimentierkompetenzen

Die Beurteilung von Schülerexperimentierleistungen beschreibt neben der Analyse und Planung von Experimentalunterricht eine von drei postulierten Teilkompetenzen des Kompetenzmodells zu *Vermittlungs- und Beurteilungskompetenzen zum Experimentieren* (Hasse, Joachim, Bögeholz & Hammann, 2014).

---

10 Der *Ingenieursmodus* bezeichnet ein Vorgehen, das darauf zielt, einen Effekt bzw. ein bestimmtes Ergebnis zu erzeugen, anstatt ursächliche Wirkungen durch eine systematische Variation von Variablen nachzuweisen.

11 Ein *confirmation bias* beschreibt das Phänomen, wenn Personen dazu neigen, ihre Erwartungen zu bestätigen. Sie ignorieren dann Ansätze bzw. Daten, die ihre Erwartungen nicht stützen. In der Folge haben sie Schwierigkeiten, Hypothesen zu revidieren.

Beurteilungen von Biologielehrkräften beziehen sich sowohl auf abschließende Benotungen von Schülerleistungen als auch auf unterrichtsbegleitende Leistungsbeurteilungen zur adaptiven Gestaltung von Lehr-Lernprozessen (vgl. Brunner, Anders, Hachfeld & Krauss, 2011; Hasselhorn & Gold, 2013). „All diese Tätigkeiten des Beurteilens und Bewertens“ werden als Diagnostizieren zusammengefasst (Hasselhorn & Gold, 2013, S. 389).

Damit Schülerleistungen transparent und gerecht beurteilt werden können, sollte die Diagnose anhand eines Kategorien- bzw. Klassifikationssystems erfolgen (ebd.). Häufig finden Beurteilungen im Unterrichtsalltag jedoch „implizit und unreflektiert auf der Grundlage subjektiver Theorien anstelle theoretischen Wissens“ statt (Hesse, 2014, S. 16).

Diagnostische Kompetenz umfasst „die Fähigkeiten und Fertigkeiten der angehenden Lehrkräfte […] Lernvoraussetzungen, Lernergebnisse sowie Lernprozesse von Schülern angemessen zu erfassen und einzuschätzen. Zudem ist auch das Erkennen von fachspezifischen Interessen und Motiven Bestandteil von diagnostischer Kompetenz. Weiterhin gehören auch Wissensbestandteile über typische Befundlagen (z. B. Fehlvorstellungen), für die Diagnostik nutzbare Instrumente aber auch Kenntnisse zu idealtypischen Entwicklungsverläufen, um zu diagnostizieren, ob und in welchem Maße eine Abweichung vorliegt, dazu“ (Cappell, 2013, S. 18).

Neben der Fähigkeit zur Beurteilung von Merkmalen von Lernenden wird die Fähigkeit zur Einschätzung von Lern- und Aufgabenanforderungen als ein Indikator diagnostischer Kompetenz betont (Artelt & Gräsel, 2009; Dübbelde, 2013; Schrader, 2008).

Diagnostik wird von Hesse (2014, S. 16) „für das Erreichen von kontinuierlichen Lernfortschritten im Unterricht und für gezielte und effiziente Förderung [für] unverzichtbar“ gehalten (vgl. auch Schrader, 2008). Entsprechend verstehen Artelt und Gräsel „diagnostische Kompetenz“ als Schlüsselkompetenz in Lehr- und Lernzusammenhängen (2009, S. 157). Diagnose und Förderung werden daher begründet durch die Standards der Lehrerbildung eingefordert (KMK, 2004a).

Insgesamt ist festzuhalten, dass die Begrifflichkeiten Beurteilen/Beurteilung/Beurteilungskompetenz und Diagnostizieren/Diagnose/diagnostische Kompetenz nicht einheitlich in der Literatur verwendet werden. Während z. B. Hasselhorn und Gold (2013) hier differenzieren, kommt Schrader zu dem Schluss: „Diagnostische Kompetenz wird häufig mit Beurteilungskompetenz gleichgesetzt“ (2009, S. 237).

Seitens der Biologiedidaktik wurden diagnostische Kompetenzen von Gießener Biologielehramtsstudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung mittels *paper-pencil* Tests und Videovignetten untersucht (Dübbelde, 2013). Dabei wurden Studierende u. a. gebeten, Schülerleistungen vorgegebenen Niveaus zuzuordnen bzw. Schülerleistungen im Hinblick auf gegebene Beurteilungskriterien zu überprüfen. Insgesamt betrachtet Dübbelde (2013) sowohl Status- als auch Prozessdiagnostik[12] in einem breiten Ansatz. In Bezug auf Reliabilität (Cronbachs $\alpha = .55$[13] für kombiniertes Status- und Prozessdiagnostikinstrument von 36 Items bei $N = 57$, ebd., S. 189) und Validität (u. a. konvergente und diskriminante Validierung) besteht jedoch weiterer Forschungsbedarf. Dübbelde (2013) diskutiert eine mögliche Mehrdimensionalität des Konstruktes Diagnosekompetenz als Begründung für die kaum zufriedenstellenden Reliabilitäten. Ungeklärt bleibt zudem die Generalisierbarkeit der Ergebnisse (Studierende einer Hochschule) und inwiefern Lehramtsstudierende eigenständig sachgerecht Beurteilungskriterien für Schülerexperimentierkompetenzen anlegen können.

5.2.3 Beschreibung von Beurteilungskompetenz für Experimentierkompetenzen

Zunächst wird der Kompetenzbegriff spezifiziert, der der vorliegenden Forschung zugrunde liegt. Mit Kompetenzen werden im Folgenden „context-specific cognitive dispositions that are acquired by learning and needed to successfully cope with certain situations or tasks in specific domains“ (Klieme, Hartig & Rauch, 2008, S. 9) bezeichnet. Fokussiert wird die kognitive Leistungsdisposition, die erforderlich ist, um experimentelle Schülerkompetenzen im Biologieunterricht fachlich adäquat beurteilen zu können. Als ein Kern von Beurteilungskompetenz für Schülerexperimentierkompetenzen wird in der Folge die Fähigkeit zur integrierten Anwendung fachdidaktischen Wissens gefasst. Zum fachdidaktischen Wissen zählt z. B. *Wissen über fachbezogene Schülerkognitionen* und *Wissen über die Anforderungen von unterrichtsfachbezogenen Aufgaben* (Brunner et al., 2011, S. 216). Da es um die Beurteilung von Lernvoraussetzungen bzw. Lernleistungen bezogen auf Experimentierkompetenzen geht, ist die Fähigkeit zum Bezug von Schülerleistungen auf fachmethodisches Wissen über Erkenntnisgewinnung in biologischen Kontexten entscheidend. Ein vertieftes Verständnis der curricularen Inhalte des jeweiligen Faches zählt dabei zum Fachwissen (ebd.).

---

12 Statusdiagnostik dient zur Erfassung von Kompetenzen, um Lernvoraussetzungen, Lernzwischenstände oder Lernergebnisse nach Interventionen zu bestimmen, während bei Prozessdiagnostik Prozesse einzelner Aufgabenbearbeitungen im Zentrum stehen (Dübbelde, 2013, S. 22f.).

13 für Statusdiagnostik-Teil mit 17 Items: $\alpha = .50$ und für Prozessdiagnostik-Teil mit 19 Items: $\alpha = .34$

Aus normativer Sicht (siehe 5.2.1), aus fachmethodischer Perspektive sowie auf Grundlage der Schülervorstellungs- und Schülerkompetenzforschung sind die drei Phasen des Experimentierens[14] zentrale Bezugspunkte bei einer Beurteilung von Schülerexperimentierkompetenzen (KMK, 2004b; Hammann, 2004).

Im Folgenden werden Beurteilungsanforderungen für (angehende) Biologielehrkräfte konkretisiert: Eine Beurteilung der Experimentierkompetenzen im Bereich Hypothesenbildung erfordert eine Prüfung, inwiefern Vermutungen von Lernenden fachlich begründet werden, inwiefern die formulierten Schülerhypothesen empirisch überprüfbar sind, inwiefern sie inhaltlich zur Fragestellung passen und inwiefern mehrere mögliche Hypothesen für die Erklärung eines Problems herangezogen werden (Hammann, 2004; Klahr, Fay & Dunbar, 1993; Schmiemann & Mayer, 2013; vgl. Klahr, 2000). Für die Diagnose der Fähigkeiten zur Planung von Experimenten sind die folgenden Kriterien beurteilungsrelevant: systematische Variation der Testvariablen bei Konstanthalten der Kontrollvariablen, Berücksichtigung von Kontrollansätzen und präzise Dokumentation des Experimentes zur Sicherung der Durchführungsobjektivität (Hammann, 2004; Chen & Klahr, 1999; Mayer & Ziemek, 2006; Schauble, 1996). Weiterhin ist beurteilungsrelevant, inwiefern Schülerinnen und Schüler ihre Experimente so planen, dass ursächliche Wirkungen mit dem Experiment nachgewiesen werden können – anstelle z. B. im *Ingenieursmodus* zu planen. Bei der Beurteilung der Fähigkeiten von Lernenden zur Auswertung der Daten gilt es zu prüfen, inwiefern eine Fehleranalyse durchgeführt wird, inwiefern in der Auswertung Bezug auf die Hypothese genommen wird und inwiefern den Schülerauswertungen ein möglicher *confirmation bias* zugrunde liegt bzw. ob die gezogenen Schlussfolgerungen logisch auf Basis der vorliegenden Evidenzen erfolgen (Dübbelde, 2013; Hammann, 2004; Klahr, 2000; Mayer & Ziemek, 2006).

Kriterien zur Beurteilung von Schülerleistungen sollen auf rationalen Gründen basieren und fachlich definierte Leistungserwartungen reflektieren (vgl. sachliche Bezugsnorm in Hasselhorn & Gold, 2013). Bei der Beurteilung der Leistungen soll einbezogen werden, inwiefern Diskrepanzen von Schülervorstellungen zu wissenschaftlichen Konzepten vorliegen (Cappell, 2013; vgl. Hammann, Phan, Ehmer & Bayrhuber, 2006). Bei Beurteilungskompetenz geht es letztendlich um die Fähigkeit zur Bewältigung von Beurteilungsanforderungen (s. o.). Über die Erfassung von Lernvoraussetzungen und -leistungen zur

[14] Das Experimentieren enthält zudem die Entwicklung von Fragestellungen. Im SDDS-Modell wird mit vorgegebenen Fragestellungen gearbeitet. Experimentieren wird als Problemlösen verstanden. Das Problemlösen wird als *Hypothesen bilden*, *Experimente entwickeln* und *Daten analysieren* verstanden.

Beurteilungskompetenz können Informationen für gezielte Förderung bereitgestellt werden (Hesse, 2014). Die Forschungsfrage des vorliegenden Beitrages lautet daher:

- Wie kann Beurteilungskompetenz von Biologielehramtsstudierenden für Experimentierkompetenzen von Schülerinnen und Schülern empirisch erfasst werden?

Wissen über Schülerkompetenzen und Schülervorstellungen zum Experimentieren ist Inhalt moderner biologiedidaktischer Veranstaltungen im Lehramtsstudium. Wissen über deren Beurteilung bzw. Diagnose und die schulcurriculare Relevanz der Erkenntnisgewinnung mittels des Experimentierens (siehe 5.2.1) kann im Biologielehramtsstudium ebenfalls erworben werden. Auf diesen Überlegungen basiert die Annahme, dass Beurteilungskompetenz im Verlauf des Lehramtsstudiums erworben werden kann. Daher kann vermutet werden, dass Biologielehramtsstudierende höherer Semester bzw. späterer Studienabschnitte über größere Beurteilungskompetenz für Schülerexperimentierkompetenzen verfügen als Studierende in weniger fortgeschrittenen Ausbildungsphasen.

5.3 Methodische Anlage

5.3.1 Stichprobe

Die hier vorgestellte Studie (01-07/2014) erfolgte mit 145 Biologielehramtsstudierenden (davon 78.5 % weiblich) von acht bundesdeutschen Universitäten aus vier Bundesländern[15]. Mit mindestens 15 Studierenden pro Standort haben die Universitäten Münster (*n* = 42 [24 Bachelor-Studierende/18 Master-Studierende]), Göttingen (*n* = 30 [7/23]), Köln (*n* = 27[16] [26/0]) und Rostock (*n* = 15[16] [1/0/12 Lehramt auf Staatsexamen]) teilgenommen. Unter den Testpersonen befinden sich 49.3 % Bachelorstudierende, 33.8 % Masterstudierende und 16.9 % Studierende, die ein Staatsexamen anstreben[17]. Die Stichprobe umfasst Lehramtsstudierende für unterschiedliche Schulformen (Gymnasium und Gesamtschule mit Sekundarstufe I/II: 50.3 %, Haupt-, Real-, Regional- und Gesamtschule mit Sekundarstufe I: 35.8 %, Sonderpädagogik: 7.6 %, Berufskolleg: 4.1 % und Grundschule: 2.1 %[18]).

15 Nordrhein-Westfalen: Münster, Köln; Niedersachsen: Göttingen, Hannover; Mecklenburg-Vorpommern: Rostock; Bayern: Bamberg, Erlangen, München.

16 Angaben zur Art des Biologielehramtsstudiums fehlen bei einigen Studierenden.

17 Die Berechnungen erfolgten mit den Personen, die Angaben zur Art ihres Biologielehramtsstudiums machten.

18 Alle drei Studierenden für das Grundschullehramt studieren Didaktik der Biologie.

5.3.2 Testheftdesign und Aufgabenzusammenstellung

Die Testhefte enthielten zu Beginn Fragen zu personenbezogenen Daten wie Geschlecht, Alter, Studiengang, studierte Semester und Studienfächer. Darauf folgte eine Beschreibung zum Aufbau des Testheftes und der Aufgaben. Daran anschließend wurden neun verschiedene Aufgaben zur Beurteilung von Schülerleistungen präsentiert.

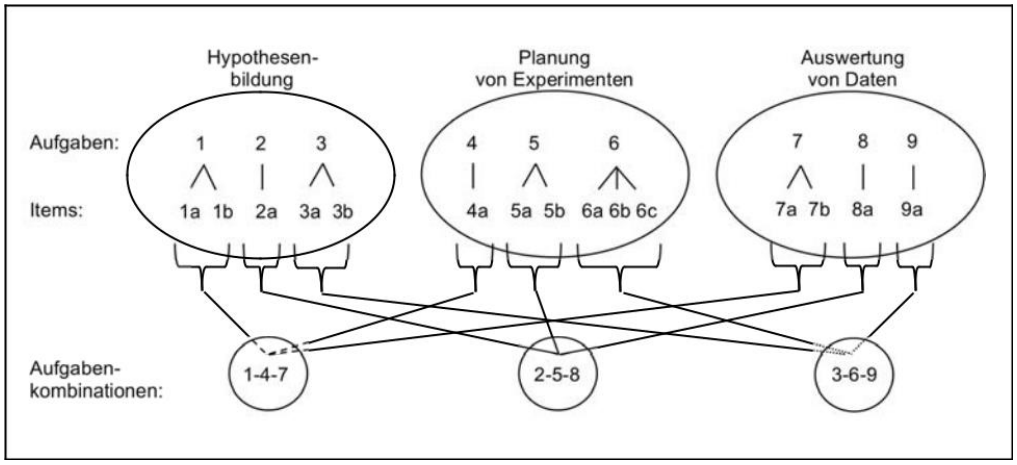


Abbildung 5-1 Erfassung von Beurteilungskompetenz für Experimentierkompetenzen: Überblick über Aufgaben, gescorte Items und Aufgabenkombinationen in den Testheften.

Insgesamt wurde ein Aufgabenpool entwickelt, der Bearbeitungskontexte für die Sekundarstufe I abdeckt. Für die Klassenstufen 5/6 wurde die Samenkeimung als exemplarischer Kontext ausgewählt, für die Klassenstufen 7/8 die Fotosynthese und für die Klassenstufen 9/10 die Enzymatik. Für jeden Bearbeitungskontext wurden neun unterschiedliche Aufgaben entwickelt. Sie sind strukturgleich für alle Kontexte.

Pro Kontext fokussieren jeweils drei der neun Aufgaben die Fähigkeit von Lernenden zur Beurteilung der Hypothesenbildung, die Fähigkeit zur Beurteilung der Experimentplanung und die Fähigkeit zur Beurteilung der Datenauswertung (Abbildung 5-1, Appendix).

Die drei Aufgaben pro Phase des Experimentierens greifen einander komplementäre Beurteilungsanforderungen auf. Ziel war es, alle resultierenden 27 Aufgaben mit vertretbarer Belastung der Testpersonen zu erproben. Dazu wurde ein *incomplete block design* verwendet (Tabelle 5-1; Frey, Hartig & Rupp, 2009).

Tabelle 5-1 Testheftdesign der Pilotstudie

Position	TH 1	TH 2	TH 3	TH 4	TH 5	TH 6	TH 7	TH 8	TH 9
1	$S_{1\text{-}4\text{-}7}$	$S_{2\text{-}5\text{-}8}$	$S_{3\text{-}6\text{-}9}$	$F_{1\text{-}4\text{-}7}$	$F_{2\text{-}5\text{-}8}$	$F_{3\text{-}6\text{-}9}$	$E_{1\text{-}4\text{-}7}$	$E_{2\text{-}5\text{-}8}$	$E_{3\text{-}6\text{-}9}$
2	$F_{2\text{-}5\text{-}8}$	$F_{3\text{-}6\text{-}9}$	$E_{1\text{-}4\text{-}7}$	$E_{2\text{-}5\text{-}8}$	$E_{3\text{-}6\text{-}9}$	$S_{1\text{-}4\text{-}7}$	$S_{2\text{-}5\text{-}8}$	$S_{3\text{-}6\text{-}9}$	$F_{1\text{-}4\text{-}7}$
3	$E_{3\text{-}6\text{-}9}$	$E_{1\text{-}4\text{-}7}$	$F_{2\text{-}5\text{-}8}$	$S_{3\text{-}6\text{-}9}$	$S_{1\text{-}4\text{-}7}$	$E_{2\text{-}5\text{-}8}$	$F_{3\text{-}6\text{-}9}$	$F_{1\text{-}4\text{-}7}$	$S_{2\text{-}5\text{-}8}$

Anmerkungen: TH = Testheft; S = Samenkeimung, F = Fotosynthese, E = Enzymatik; tiefgestellte Ziffern stehen für Aufgabenkombinationen vgl. Abbildung 5-1.

Zusammengestellt wurden neun Testhefte mit je neun unterschiedlichen Aufgaben (Tabelle 5-1). Jedes Testheft (TH) enthält drei Aufgaben zu jedem der drei Bearbeitungskontexte (z. B. TH 1 in Tabelle 5-1). Die drei Bearbeitungskontexte stellen die drei Blöcke Samenkeimung (S), Fotosynthese (F) und Enzymatik (E) dar. Jeder Block fasst je eine Aufgabe zur Hypothesenbildung (1, 2 oder 3 in erster Position), eine Aufgabe zur Planung von Experimenten (4, 5 oder 6 in zweiter Position) und eine Aufgabe zur Auswertung von Daten (7, 8 oder 9 in dritter Position). Die Zusammenstellung der Aufgabenblöcke zu jedem Kontext erfolgt mittels der folgenden drei Aufgabenkombinationen: 1-4-7, 2-5-8 und 3-6-9 (Tabelle 5-1, Abbildung 5-1). Mit dem Testheftdesign kommen alle drei Aufgabenkombinationen in jedem Testheft vor (Tabelle 5-1). Ziel war es, die Messung von Beurteilungskompetenz bezogen auf Schülerleistungen zu allen drei Phasen des Experimentierens nicht mit Bearbeitungskontexten oder konkreten Aufgaben(stellungen) zu konfundieren.

Die Testhefte wurden der Reihe nach ausgegeben (TH 1 bis TH 9, dann wieder beginnend mit TH 1). Die Bearbeitungszeit betrug 90 Minuten – mit der Ausnahme von zwei Standorten, wo entweder nur 60 oder nur 70 Minuten zur Verfügung standen. Bei reduzierter Testzeit wurden die Aufgaben in letzter Position nicht bearbeitet.

5.3.3 Operationalisierung von Beurteilungskompetenz und Scoring von Beurteilungsleistungen

Zum Einsatz kommt ein systematischer Aufgabenentwicklungs- und Aufgabenauswertungsansatz, der über mehrere Schritte erarbeitet wurde (Studie Lauten Denkens: N = 16, Pre-Pilotierung: N = 24, Expertenbefragung: N = 6).

Die Aufgaben beschreiben realitätsnahe Unterrichtssituationen und Schülerleistungen beim Experimentieren im Biologieunterricht (Abbildung 5-2). In der Folge werden die Lehramtsstudierenden gebeten, die präsentierte(n) Schülerleistung(en) zu beurteilen (ebd.). Entsprechend wurden Arbeitsaufträge zur Beurteilung formuliert (Appendix).

Die Schülerinnen und Schüler einer sechsten Klasse haben im Winter anhand der Wärmeisolation durch Federn das Experimentieren kennengelernt. Im Mai soll nun das Experimentieren am Thema Samenkeimung vertieft werden. In der letzten Stunde wurde die Fragestellung erarbeitet: „Welche Faktoren sind entscheidend für die Samenkeimung bei Buschbohnen?“ Dann wurden entsprechende Hypothesen aufgestellt.
In dieser Stunde sollen die Lernenden ein Experiment zur Samenkeimung von Buschbohnen planen. Paul entscheidet sich, den Einfluss der zwei Faktoren Licht und Temperatur zu untersuchen.

Paul überlegt sich ein Experiment, mit dem er herausfinden kann, ob Buschbohnen zum Keimen Licht und Wärme benötigen.

Planung und Durchführung von Paul: *Ich lege Bohnensamen in zwei Töpfe mit Erde und gieße beide Töpfe ausreichend.*

- *Topf 1 stelle ich in einen hellen Raum mit 22°C.*
- *Topf 2 stelle ich in einen Kühlschrank ohne Licht bei 4°C.*

Nach 7 Tagen macht er die Beobachtung: *Nur die Samen in Topf 1 haben gekeimt, nicht die Samen in Topf 2.*

➢ Beurteilen Sie Pauls Planung des Experimentes. Begründen Sie.

Item	Scoring
Fähigkeit zur Beurteilung von Schülerleistung mit unsystematischer Variablen-variation	(Angehende) Biologielehrkraft … • erläutert allgemein, dass mit diesem Experiment keine Aussage über den Einfluss der einzelnen Faktoren getroffen werden kann oder erläutert, dass die Variablen unsystematisch variiert wurden [Score 1]. • erläutert den unsystematischen Umgang mit Variablen am konkreten Beispiel der Samenkeimung und den Faktoren Licht und Temperatur [Score 2].

Abbildung 5-2 Aufgabenbeispiel mit Textvignette, Arbeitsauftrag zur Beurteilung und Scoring eines Items (gekürzte Fassung von Aufgabe 4 im Bearbeitungskontext Samenkeimung).

Ein Beispiel für eine typische Aufgabe wird in Abbildung 5-2 präsentiert. Es handelt sich um Aufgabe 4 umgesetzt im Kontext Samenkeimung. Bei Aufgabe 4 wird in allen Bearbeitungskontexten je ein Item gescort (0-1-2 Scoring; Abbildung 5-2). In der vorliegenden Studie stellt ein Item einen relevanten Aspekt von Beurteilungskompetenz dar und entspricht einer Auswertungseinheit für die untersuchte Kompetenz. Mit Blick auf das Item *Fähigkeit zur Beurteilung von Schülerleistung mit unsystematischer Variablenvariation* (Abbildung 5-2, Item 4a in Appendix) wurde für die Studierendenantwort „Paul berücksichtigt nicht, dass er die unabhängigen Variablen systematisch verändern muss […]“ (TH 5.15: EL18OT) *partial credit* vergeben, während die Antwort „Pauls Experiment kann die Fragestellung nicht beantworten, weil zwei Variablen vorhanden sind, die auch gleichzeitig variiert werden. Das „Nicht-Keimen“ könnte jetzt sowohl auf die niedrigere Temperatur als auch auf das fehlende Licht zurückgeführt werden. Es darf immer nur

eine Variable (unabhängige Variable) variiert werden, während eine abhängige Variable (Keimen oder nicht) untersucht wird“ (TH 1.10:CH20WI) mit *full credit* gescort wurde.

Durch den Arbeitsauftrag zur Beurteilung soll spezifisch Pauls Leistung bei der Planung des Experimentes evaluiert werden. Dabei wird um eine Begründung für die Beurteilung gebeten. Ziel ist es, herauszufinden, inwiefern erlernte, wissensbasierte Konzepte auf gegebene Experimentierleistungen in kontextualisierten Beurteilungssituationen angewendet werden können.

Während bei Aufgabe 4 nur ein Item gescort wird, werden bei fünf der neun strukturell unterschiedlichen Aufgaben je zwei oder drei Items gescort (Abbildung 5-1, Appendix). In der Regel erfolgt ein trichotomes Scoring (0-1-2). Lediglich die Items 6c und 9a werden dichotom gescort. Die Bearbeitung einer Aufgabe mit mehreren Beurteilungsanforderungen (Items) in unterschiedlicher Bearbeitungsqualität (Qualitätswerte z.B. 0-1-2) kann Rückschlüsse auf die untersuchte Beurteilungskompetenz ermöglichen.

Alle gescorten Items für die neun Aufgaben und die entsprechenden Arbeitsaufträge zur Beurteilung sind im Appendix aufgeführt. Sowohl die Arbeitsaufträge als auch die Items sind für jede Aufgabe in den drei Kontexten parallel angelegt (strukturgleiche Aufgaben).

Alle Aufgaben wurden nach einem *scoring guide* von zwei unabhängigen Raterinnen gescort. Bei abweichenden Scores wurde überwiegend argumentativ Konsens erzielt. In der Folge wurde Cohens Kappa berechnet (.97). Die Anlage des Scorings folgte schwerpunktmäßig den folgenden Prinzipien: Höhere Scores werden vergeben bei steigendem Systematisierungsgrad (vgl. z. B. Hammann, 2004) und steigendem Elaborationsgrad (vgl. z. B. Upmeier zu Belzen & Krüger, 2010). Differenziert und kontextualisiert kommentierte Schülerleistungen, kontrastiert mit definierten Leistungserwartungen, dokumentieren transparent eine Beurteilung der tatsächlichen Schülerleistung (vgl. Hasselhorn & Gold, 2013) bzw. eine Diagnose, inwiefern eine Abweichung vorliegt (vgl. Cappell, 2013). Dies greift die Annahme auf, dass eine alleinige Nennung abstrakt gefasster Konzepte wie z. B. des *Ingenieursmodus* oder des *confirmation bias* als Schülerfehler beim Experimentieren nicht deren Verständnis belegen. Diese Konzepte sind – zumindest an einigen Standorten – Gegenstand des biologiedidaktischen Ausbildungskanons. Eine kontextualisierte Erklärung von Schülerleistungen in Verbindung mit dem entsprechendem Konzept setzt hingegen ein tieferes Konzeptverständnis voraus (variabler Transfer des Konzeptes in andere Anwendungskontexte vs. Reproduktion von Wissen).

5.4 Ergebnisse

Im Folgenden werden Reliabilitäts- und Validitätsanalysen beschrieben. Datenbasis für die Reliabilitätsanalysen sind die 15 im Appendix abgedruckten Items – allerdings in unterschiedlichen Variationen von Bearbeitungskontexten (TH 1-9 in Tabelle 5-1). Von den ursprünglich 15 Items pro Testheft mit neun Aufgaben (siehe Abbildung 5-1) wurden diejenigen ausgeschlossen, die über Lösungswahrscheinlichkeiten von < 10 % und > 90 % verfügen. Prämisse für die Skalenbildung war, dass die letztendliche Skala für jedes Testheft Items aus jedem Bearbeitungskontext und jeder Phase des Experimentierens berücksichtigt. Alle zur Reliabilitätsberechnung verwendeten Items verfügen über positive Trennschärfen.

Insgesamt wurde jedes Testheft von 15 bis 17 Studierenden bearbeitet. Ausgewertet wurden nur die Personen, die alle neun Aufgaben bearbeitet haben (siehe Tabelle 5-2).

Tabelle 5-2 Reliabilitäten für die Erfassung von Beurteilungskompetenz berechnet für jedes Testheft

Testheft (*n*)	TH 1 (12)	TH 2 (13)	TH 3 (13)	TH 4 (12)	TH 5 (13)	TH 6 (12)	TH 7 (13)	TH 8 (13)	TH 9 (12)
Anzahl Items	9	6	6	8	10	10	8	6	6
Cronbachs α	.69	.73	.71	.74	.70	.72	.74	.57	.75
Spearman-Brown bei 10 Items	.71	.82	.80	.78	.70	.72	.78	.69	.83

Anmerkungen: TH = Testheft; *n* = Anzahl der ausgewerteten Personen.

Die Reliabilitäten für sieben von neun Testheften liegen bei mindestens .70 (Cronbachs α siehe Tabelle 5-2). Für die Testhefte wurden in die gebildeten Skalen zwischen sechs und zehn Items einbezogen. Tabelle 5-2 gibt zudem für alle Testhefte untereinander vergleichbare extrapolierte Cronbachs αs (nach Spearman-Brown) an, wobei stets zehn Items pro Testheft zugrunde gelegt wurden. Mit Ausnahme von Testheft 8 liegen die geschätzten Reliabilitäten bei mindestens .70; bei fünf der neun Testhefte liegen die geschätzten Reliabilitäten zwischen .78 und .83.

Aufgrund der geringen Testpersonenzahlen pro Testheft wurden Testhefte mit identischen Aufgaben und Items zusammen analysiert. Eine gemeinsame Analyse der Testhefte 1 und 5 ergab einen Cronbachs α von .58 für neun Items bei $n = 25$. Der Cronbachs α Wert für die gemeinsame Analyse der Testhefte 2 und 7 ist für $n = 26$: .67 bei sechs Items.

Eine Überprüfung der Annahme, dass Masterstudierende über bessere Beurteilungskompetenz verfügen als Bachelorstudierende, ergab folgendes Bild: Bei der kombinierten Analyse der Testhefte 2 und 7 zeigten Studierende aus dem Master (M = .58[19], SD = .28, n = 12) eine höhere Beurteilungskompetenz als Bachelorstudierende (M = .25, SD = .17, n = 14). Die Mittelwertunterschiede sind signifikant ($t_{(24)}$ = -3.66, p < .001). Die gemeinsame Analyse der Testhefte 1 und 5 hingegen ergab keine nachweisbaren Unterschiede ($t_{(25)}$ = -.81, *n.s.*, Masterstudierende: M = .41, SD = .17, n = 9; Bachelorstudierende: M = .35, *SD:* .19, n = 18). Regressionsanalytisch zeigte sich ein vergleichbares Befundmuster für den Einfluss von studierten Semestern auf die Beurteilungskompetenz: Während für die Studierenden, die die Testhefte 2 und 7 bearbeitet hatten, ein starker Einfluss der Anzahl der studierten Semester nachweisbar war (r^2= .41, p < .001), zeigte sich kein signifikanter Effekt für die Studierenden, die die Testhefte 1 und 5 bearbeiteten.

5.5 Zusammenfassung, Diskussion und Ausblick

5.5.1 Erkenntnisse und Potenziale der Pilotstudie

Zentraler Bestandteil der Beurteilungskompetenz von (angehenden) Biologielehrkräften für Schülerexperimentierkompetenzen ist die Bewältigung von Beurteilungsanforderungen unter Verständnis von alternativen Schülervorgehensweisen, z. B. durch Erkennen von typischen Schülervorstellungen beim Experimentieren (vgl. Hammann, 2004). Letzteres ist eine Voraussetzung, um „den Lehr-Lern-Prozess gezielt zu […] gestalten und das Lernen der Schüler zu optimieren“ (Schrader, 2008, S. 168).

Präsentiert wird im Artikel ein Ansatz zur Erfassung von Beurteilungskompetenz anhand eines systematischen Aufgabenentwicklungs- und Aufgabenauswertungsansatzes. Der Ansatz wurde exemplarisch für den Biologieunterricht der Sekundarstufe I konzipiert. Der Vorteil dieser Herangehensweise ist die Passung von zu vermittelnden Schülerkompetenzen (z. B. systematischer Umgang mit Variablen) und dem Beurteilen eben dieser transferfähigen Kompetenzen. Der Ansatz kann für Experimentalunterricht in der Sekundarstufe II im Fach Biologie weiterentwickelt und für weitere naturwissenschaftliche Experimentalfächer angepasst werden.

Die hier referierte Studie ergab mit offenen Aufgaben akzeptable Reliabilitäten (Tabelle 5-2), die deutlich über den Werten liegen, die bisher mittels geschlossener Aufgaben zu

[19] Vor den Datenanalysen wurden alle trichotom gescorten Items von 0-1-2 auf 0.5-1 umcodiert, um sie gemeinsam mit den dichotomen (0-1 codierten) Items zu analysieren.

diagnostischen Kompetenzen von Biologielehramtsstudierenden berichtet wurden (maximal .55 in Dübbelde, 2013, S. 189). Damit kann die Hypothese aufrechterhalten werden, dass es sich bei Beurteilungskompetenz zum Experimentieren um eine eindimensionale Skala handelt, die sich auf alle drei Phasen des Experimentierens bezieht. Bei diagnostischer Kompetenz wird zwar über Mehrdimensionalität diskutiert (z. B. Cappell, 2013; Dübbelde, 2013)[20], das zu erfassende Konstrukt ist dort aber weiter gefasst als die hier im Fokus stehende Beurteilungskompetenz.

Weiterhin erfolgten erste Analysen hinsichtlich der Validität von Beurteilungskompetenz. Dazu wurden Master- und Bachelorstudierende verglichen. Masterstudierende haben i. d. R. mehr fachdidaktische Lerngelegenheiten in ihrer Ausbildung wahrgenommen als Bachelorstudierende – betrachtet man jeden einzelnen universitären Ausbildungsgang gesondert. Trotz nicht eindeutiger Befundlage deutet die Analyse der Testhefte 2 und 7 darauf hin, dass das gemessene latente Konstrukt potenziell erlernbar ist, womit ein Hinweis auf eine Kompetenz vorläge (vgl. z. B. Klieme et al., 2008). Dieser Hinweis ist sehr begrenzt, da er nur in einer von zwei Testheftkombinationen auftritt. Wohl aber wird der Teilbefund für die Kombination der Testhefte 2 und 7 durch einen zweiten Befund flankiert: Regressionsanalytisch zeigte sich, dass die Anzahl studierter Semester einen starken Einfluss auf die Beurteilungskompetenz hat. Dennoch ist Vorsicht geboten, die (Teil-)Befunde dieser kleinen Studie nicht zu überschätzen. Zum einen ist die Stichprobe sehr begrenzt und zum anderen handelt es sich um querschnittlich erhobene Daten.

Die vorliegende Studie stellt einen großen Aufgabenpool für die folgende Hauptstudie zur Kompetenzmodellierung und -messung bereit. Dabei fasst die referierte Studie eine Reihe strukturgleicher Aufgaben zur Erfassung von Beurteilungskompetenz in unterschiedlichen Bearbeitungskontexten der Sekundarstufe I. Dieser Pool ist eine wertvolle Vorarbeit für Veränderungsmessungen – beispielsweise für Prä-Post-Kontrollgruppendesigns (Interventionsstudien zur Kompetenzförderung) und Längsschnittstudien.

5.5.2 Grenzen der Pilotstudie

Ziel war es, auf möglichst systematische Art und Weise Aufgaben mit Beurteilungsanforderungen in verschiedenen Bearbeitungskontexten zu erproben. Damit unmittelbar verbunden ist eine begrenzte Aussagekraft der Testergebnisse. Gründe dafür sind die

[20] Vgl. auch Auffassung von diagnostischen Fähigkeiten als „mehrdimensionale Kompetenzfacette, die eine Integration mehrerer Kompetenzfacetten des fachdidaktischen und pädagogischen Wissens erfordert“ in Brunner et al. (2011, S. 217).

geringe Anzahl der Personen, die das gleiche Testheft bearbeitet haben, der vergleichsweise große Anteil an Bachelorstudierenden in der Stichprobe sowie der unterschiedliche Anteil von Studierenden pro Studienabschnitt an den verschiedenen universitären Standorten (Übergewicht an Masterstudierenden in Göttingen und an Bachelorstudierenden in Köln). Zudem verfügen die einzelnen Standorte über unterschiedliche Ausbildungscurricula und Ausbildungspraktiken. Somit ist die Aussagekraft des Vergleichs von Master- und Bachelorstudierenden – wie auch der Analysen aufgrund studierter Semester – begrenzt.

Ein weiterer limitierender Faktor ist die zumutbare Testzeit – und damit einhergehend der Feldzugang. Ein Instrument mit offenen Aufgaben, das Beurteilungskompetenz für Schülerexperimentierkompetenzen curricular valide für die Sekundarstufe I abdeckt, erfordert die Messung von genügend Indikatoren – und damit eine gewisse Testzeit. Die Testzeit und die Anzahl rekrutierbarer Personen sind starke Limitationen für Studien zur Messung von Lehrerkompetenzen. Das Problem kann über Multi-Matrix Ansätze auch nicht gänzlich aufgefangen werden.

Eine offene Frage besteht darin, inwiefern unser gewählter Scoring-Ansatz, der höhere Scores für die kontextualisierte Anwendung abstrakter Konzepte bzw. Prinzipien vergibt als für prinzipienbasierte Begründungen, für Lehramtsstudierende in IRT-Modellierungen trägt. Vorstellbar wäre, dass der Ansatz zur Erfassung von Beurteilungskompetenz für Studierende und ggf. auch für Referendarinnen und Referendare geeignet ist. Aufgrund der Expertiseforschung, die zeigt, dass mit höherer Expertise Probleme prinzipienbasiert und weniger konkret analysiert werden (vgl. z. B. Chi, Feltovich & Glaser, 1981), wäre es aber denkbar, dass Beurteilungskompetenz von erfahrenen Lehrkräften anders erfasst werden müsste. Durch Studium, Referendariat und Lehrerfortbildungen – sowie durch Erfahrungen in der Anwendung der gelernten Konzepte – kann ein möglicher Bias bedingt sein (vgl. *differential item functionning* beim Vergleich von Befragten vor und nach Interventionen in Boone, Staver & Yale, 2014). Lehramtsstudierende können im Laufe ihres Berufslebens einen gewissen Expertisegrad im Beurteilen von Experimentierkompetenzen erwerben. Eine virulente Frage ist daher, inwiefern „does our item-defined meterstick“ operate in the same way for different groups of respondents?“ (ebd., S. 274).

5.5.3 Ausblick

Die Arbeiten zum Kompetenzmodell zu *Vermittlungs- und Beurteilungskompetenzen zum Experimentieren* bedürfen weiterer Fundierung. Derzeit erfolgen die Auswertungen der Hauptstudiendaten (ca. *N* = 500 von 18 Universitäten aus sieben Bundesländern).

Durch den Einsatz eines Testheftes für Vermittlungskompetenzen (Analyse und Planung von Experimentalunterricht, Universität Münster) nebst eines Testheftes für Beurteilungskompetenz (Universität Göttingen) werden ein- und mehrdimensionale IRT-Modellierungen möglich. Zur Validierung von Beurteilungskompetenz wurden Selbstwirksamkeitserwartungen von Biologielehramtsstudierenden, z. B. in den Dimensionen Schwierigkeiten von Schülerinnen und Schülern und Leistungsbeurteilung (Mahler, 2014), erhoben. Zudem wurden ihre diagnostischen Kompetenzen erfasst. Dafür kam eine gekürzte Variante des *Beurteilungsbogens zur Statusdiagnostik* des kombinierten Instrumentes nach Dübbelde (2013, Anhang S. 25) zum Einsatz.

Zentrale Ziele sind die Klärung der Reliabilität und Validität und letztendlich der Dimensionalität von Vermittlungs- und Beurteilungskompetenzen. Die laufende Studie verspricht belastbare empirische Fundierungen des postulierten Kompetenzmodells. Zudem sind weitere Studien angedacht, die schließlich auch eine Überprüfung des Modells mittels experimenteller Validierung ermöglichen.

Danksagung

Der ExMo-Verbund wird vom BMBF im Rahmen von KoKoHs gefördert. Dank gilt weiterhin den befragten Studierenden und den befragten Expertinnen. Darüber hinaus danken wir Claus Carstensen zwei Gutachterinnen bzw. Gutachtern für ihre hilfreichen kritisch-konstruktiven Kommentare zur Weiterentwicklung des Manuskriptes.

Literatur

Artelt, C. & Gräsel, C. (2009). Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*, 157-160.

Baumert, J. & Lehmann, R. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich – Deskriptive Befunde.* Opladen: Leske & Budrich.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences.* Dordrecht: Springer.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 213-234). Münster: Waxmann.

Cappell, J. (2013). *Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase*. Berlin: Logos.

Chen, Z. & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development, 70*, 1098-1120.

Chi, M. T. H., Feltovich, P. J. & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science, 5*, 121-152.

Dübbelde, G. (2013). *Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung.* (Dissertation, Universität Kassel). Zugriff am 19.06.2015 unter https://kobra.bibliothek.uni-kassel.de/handle/urn:nbn:de:hebis:34-2013122044701

Frey, A., Hartig, J. & Rupp, A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*, 39-53.

Gesellschaft für Fachdidaktik (2005a). *Fachdidaktische Kompetenzbereiche, Kompetenzen und Standards für die 1. Phase der Lehrerbildung (BA+MA)(Anlage 1).* Zugriff am 19.06.2015 unter http://fachdidak-tik.org/Ver%C3%B6ffentlichungen.html

Gesellschaft für Fachdidaktik (2005b). *Zuordnungstabelle von Kompetenzformulierungen (Anlage 4).* Zugriff am 19.06.2015 unter http://fachdidak-tik.org/Ver%C3%B6ffentlichungen.html

Hammann, M. (2004). Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung – dargestellt anhand von Kompetenzen beim Experimentieren. *MNU – Der mathematische und naturwissenschaftliche Unterricht*, *57*, 196-203.

Hammann, M., Phan, T. T. H., Ehmer, M. & Bayrhuber, H. (2006). Fehlerfrei Experimentieren. *MNU – Der mathematische und naturwissenschaftliche Unterricht, 59*, 292-299.

Hasse, S., Joachim, C., Bögeholz, S. & Hammann, M. (2014). Assessing Teaching and Assessment Competences of Biology Teacher Trainees: Lessons from Item Development. *The International Journal of Education in Mathematics, Science and Technology, 2*, 191-205.

Hasselhorn, M. & Gold, A. (2013). *Pädagogische Psychologie: Erfolgreiches Lernen und Lehren* (3. überarb. Aufl.). Stuttgart: W. Kohlhammer.

Hesse, I. (2014). Pädagogisch-psychologische Diagnostik für Lehrkräfte – Herausforderung, Aufgaben, Probleme. In A. Fischer, C. Hößle, S. Jahnke-Klein, H. Kiper, M. Komorek, J. Michaelis, … J. Sjuts (Hrsg.), *Diagnostik für lernwirksamen Unterricht* (S. 15-39). Baltmannsweiler: Schneider.

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

Klahr, D., Fay, A. L. & Dunbar, K. (1993). Heuristics for Scientific Experimentation: A Developmental Study. *Cognitive Psychology, 25*, 111-146.

Klieme, E., Hartig, J. & Rauch, D. (2008). The Concept of Competence in Educational Contexts. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 3-22). Göttingen: Hogrefe.

Kultusministerkonferenz (2004a). *Standards für die Lehrerbildung: Bildungswissenschaften.* Zugriff am 19.06.2015 unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004 _12_16-Standards-Lehrerbildung.pdf

Kultusministerkonferenz (2004b). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. Zugriff am 19.06.2015 unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/ 2004_12_16-Bildungsstandards-Biologie.pdf

Mahler, H. (2014). *Selbstwirksamkeitserwartungen angehender Biologielehrkräfte – Entwicklung eines Messinstrumentes.* (Masterarbeit für Master of Education, Universität Göttingen).

Mayer, J. & Ziemek, H.-P. (2006). Offenes Experimentieren: Forschendes Lernen im Biologieunterricht. *Unterricht Biologie, 317*, 2-12.

Meier, M. & Wellnitz, N. (2013). Beobachten, Vergleichen und Experimentieren mit Wasserflöhen: Biologische Erkenntnismethoden praktisch anwenden. *Praxis der Naturwissenschaften Biologie in der Schule, 62*, 4-10.

Nieders. Kultusministerium (2009). *Kerncurriculum für das Gymnasium – gymnasiale Oberstufe, die Gesamtschule – gymnasiale Oberstufe, das Fachgymnasium, das Abendgymnasium, das Kolleg: Biologie*. Zugriff am 19.06.2015 unter http://db2.nibis.de/1db/cuvo/datei/kc_biologie_go_i_2009.pdf

Schauble, L. (1996). The Development of Scientific Reasoning in Knowledge-Rich Contexts. *Developmental Psychology, 32*, 102-119.

Schmiemann, P. & Mayer, J. (Hrsg.). (2013). *Experimentieren Sie! Biologieunterricht mit Aha-Effekt. Selbständiges, kompetenzorientiertes Erarbeiten von Lehrplaninhalten*. Berlin: Cornelsen.

Schrader, F.-W. (2008). Diagnoseleistungen und diagnostische Kompetenzen von Lehrkräften. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (Bd. 10, S. 168-177). Göttingen: Hogrefe.

Schrader, F.-W. (2009). Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23,* 237-245.

Upmeier zu Belzen, A. & Krüger, D. (2010). Modellkompetenz im Biologieunterricht: Struktur und Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften, 16,* 41-57.

Appendix

Appendix: Kurzfassung der Arbeitsaufträge zur Beurteilung durch Biologielehramtsstudierende in den Aufgaben und gescorte Items

Hypothesenbildung – Aufgaben 1-3
Arbeitsaufträge für 1-3: Beurteilen Sie die Hypothese. Erläutern Sie Ihre Beurteilung.

1 • Eine Begründung für die Hypothese der Schülerin bzw. des Schülers fehlt. [Item 1a]
• Die Hypothese der Schülerin bzw. des Schülers ist testbar. [Item 1b]
2 • Die Hypothese der Schülerinnen und Schüler ist nicht testbar. [Item 2a]
3 • Die Hypothese der Schülerinnen und Schüler hat Bezug zur Frage- bzw. Problemstellung. [Item 3a]
• Die Schülerinnen und Schüler stellen nur eine von mehreren möglichen Hypothesen auf. [Item 3b]

Planung von Experimenten – Aufgaben 4-6
Arbeitsaufträge für 4-6: Beurteilen Sie die Planung des Experimentes. Begründen Sie.

4 • Die Variablenvariation (der Schülerin und) des Schülers erfolgt unsystematisch bzw. das Experiment lässt keine Aussage zum Einfluss der einzelnen Variablen zu. [Item 4a]
5 • Die Schülerin bzw. der Schüler agiert im Ingenieursmodus. [Item 5a]
• Die Schülerin bzw. der Schüler plant das Experiment so, dass ihre/seine Annahme bestätigt wird *(confirmation bias)*. [Item 5b]
6 • Kontrollansätze sind in der Experimentplanung der Schülerin vorhanden. [Item 6a]
• Die Planung der Schülerin ist ungenau und damit ist eine Wiederholbarkeit nicht gegeben. [Item 6b]
• Der Umgang mit Variablen erfolgt von der Schülerin systematisch bzw. das Experiment der Schülerin ermöglicht Aussagen über den Einfluss der einzelnen Variablen. [Item 6c]

Auswertung von Daten – Aufgaben 7-9
Arbeitsaufträge für 7-9: Beurteilen Sie die Datenauswertung. Begründen Sie.

7 • Die Datenauswertung der Schülerin bzw. des Schülers berücksichtigt nicht alle Daten. [Item 7a]
• Bestimmte Daten werden von der Schülerin bzw. dem Schüler vermutlich aufgrund von Überzeugungen ignoriert *(confirmation bias)*. [Item 7b]
8 • Abweichende Ergebnisse werden von den Schülerinnen und Schülern nicht diskutiert. [Item 8a]
9 • Die Schülerin stellt in ihrer Schlussfolgerung keinen Bezug zur Hypothese her. [Item 9a]

6. Modeling and Measuring Pre-Service Teachers' Assessment Literacy Regarding Experimentation Competences in Biology[21]

Abstract:

Assessment literacy is a crucial aspect of teachers' professional knowledge and relevant to fostering students' learning. Concerning experimentation, teachers have to be able to assess student achievement when students form hypotheses, design experiments, and analyze data. Therefore, teachers need to be familiar with criteria for experimentation as well as student conceptions of experimentation. The present study modeled and measured 495 German pre-service teachers' *knowledge of what to assess regarding experimentation competences in biology*. We applied an open-answer format for the measurement instrument. For modeling we used item response theory (IRT). We argue that *knowledge of what to assess* regarding experimentation competences is a one-dimensional construct and we provide evidence for the validity of the measurement. Furthermore, we describe qualitative findings of pre-service teachers' *knowledge of what to assess*, in particular difficulties concerning the assessment of student conceptions as well as the use of scientific terms in the assessments. We discuss the findings in terms of implications for science teacher education and further research perspectives.

Keywords: assessment literacy; teacher education; experimentation; competence; biology

6.1 Introduction

When student experimentation competences are fostered, teachers' assessment skills come into focus. *Assessment literacy* of teachers has been found to have a significant effect on students' learning [1]. Educational assessment is closely connected to instruction and takes place regularly [2] (p. 1). It is a prerequisite to planning lessons and adapting instruction to the students' needs. Moreover, "assessment information provides feedback to the student", which can enhance achievement [2] (p. 7) [3].

[21] Joachim, C., Hammann, M., Carstensen, C. H., & Bögeholz, S. (2020). Modeling and Measuring Pre-Service Teachers' Assessment Literacy Regarding Experimentation Competences in Biology. *Education Sciences, 10*(5). https://doi.org/10.3390/educsci10050140

A central learning objective in biology are experimentation competences [4,5]. Experimentation competences are acquired successively in high school. One challenge for students is to understand how new findings in biology are gained. Specifically, German students often have misconceptions regarding experimentation [6] (p. 199). Therefore, the formation of hypotheses, the design of experiments, and the analysis of data must be practiced, and mistakes should be discussed [7]. Teachers have to be able to assess students' experimentation competences and related conceptions adequately to adapt their instruction, thus enhancing students' understanding [2] (p. 10). Assessing experimentation competences requires a range of knowledge. Disciplinary knowledge (cf. content knowledge, CK), and pedagogical content knowledge (PCK) are essential for the assessments [8] (p. 156). Teacher education in universities has to establish the essential knowledge and skills regarding assessment literacy. "knowledge is converted to skills" with increasing competence [9] (p. 70) [10]. Skills do not include "mastering a technique [. . .] [without] the use of systematic knowledge" in this contribution [11] (p. 374). Assessment literacy can be expanded in the proceeding teaching practice and the teaching career [12]. Assessment literacy is "defined as a basic understanding of educational assessment and related skills to apply such knowledge to various measures of student achievement" [8] (p. 149) cf. [13].

To date, only a few research studies have been conducted on teachers' assessment literacy for inquiry concerning biology. One of the studies has analyzed pre-service teachers' *diagnostic competence* for experimentation competences via a questionnaire with a closed answer format, for which the Cronbach's alpha for the measure of *diagnostic competence* ($\alpha = 0.50$) was low [14] (pp. 67ff., p. 189). Alternatively, the present study specifically focuses on subject-specific research regarding teacher education in biology in Germany. The paper-pencil questionnaire study applies open-ended tasks presenting classroom scenarios of experimental biology lesson activities, which meet real-life demands in assessment more closely than a closed-answer format. It aims to develop a more reliable instrument to capture pre-service teachers' knowledge of assessment criteria for experimentation and their ability to apply these criteria. Our research goals are modeling and measuring pre-service biology teachers' knowledge in the area of assessment literacy. Thereby, we aim to gain qualitative insights into pre-service teachers' strengths and weaknesses in assessing experimentation competences.

6.1.1 Assessment Literacy as Part of Professional Knowledge

Knowledge of assessment is part of teachers' professional knowledge [15]. Assessment literacy encompasses four areas of knowledge: (1) knowledge of assessment purposes, (2) knowledge of what to assess, (3) knowledge of assessment strategies, and (4) knowledge of assessment interpretation and action-taking [16]. With regard to the first area of knowledge, teachers should be familiar with the aims of assessment such as "Providing data for instructors on which to base instructional decisions" [16] (p. 213). The second area of knowledge acknowledges that knowledge of what to assess is linked to "curricular goals and to values of what is important to learn and how learning occurs" [16] (p. 214), suggesting that to assess students adequately, teachers must be knowledgeable and proficient in curriculum topics and skills. In addition, knowledge of students' misconceptions is an essential component of this area of knowledge [16] (p. 216f.). The third area of knowledge, knowledge of assessment strategies, encompasses different ways that can be applied to assessment. Teachers should be familiar with strategies for formal and informal assessment. Abell and Siegel emphasize that knowledge of assessment strategies also includes "knowledge of topic-specific assessment tasks" and "knowledge of response strategies" [16] (p. 214). Finally, the knowledge of assessment interpretation and action-taking is hallmarked, for example, by being able to use assessment results to adapt instruction [16] (p. 215).

Of these four areas of knowledge, *knowledge of what to assess* is especially content specific and fundamental, highlighting the relevance in taking a closer look at pre-service teachers' *knowledge of what to assess* regarding experimentation competences in biology in the following.

6.1.2 Assessment of Students' Experimentation Competences

Teachers' *knowledge of what to assess* comprises knowledge of concepts and processes regarding experimentation that students need to acquire and an understanding of student conceptions and difficulties. We describe teachers' knowledge of concepts and processes as well as student conceptions and difficulties in the following section.

Students need to be able to apply scientific knowledge, such as knowledge of science and knowledge about science [17]. One learning objective is procedural knowledge to understand how scientific knowledge is generated. A central method in science to gain new findings is experimentation [18] (p. 15) [19] (p. 323).

Following the general model of *Scientific Discovery as Dual Search* (SDDS) [20], experimentation competences comprise the three phases: searching hypotheses, testing the hypotheses, and evaluating the evidence [21] (p. 8). Next, we summarize requirements regarding the core facets of the three phases.

Experiments serve to examine causal relationships. Hypotheses state assumed relationships between independent and dependent variables [22] (p. 45f.). Hypotheses should be "fully specified and testable" [21] (p. 8). Hypotheses can be theoretically founded based on previous knowledge [23] (p. 9f.). Research has shown that it can be difficult for younger students to think of different explanations for a phenomenon and generate alternative hypotheses. Often, the formation of hypotheses is incomprehensive [24] (p. 245) [7] (p. 298).

For testing hypotheses, it is essential to design structured experiments and vary the independent variables (the potential causes) systematically. All other variables must be kept constant to achieve unambiguous results [18] (p. 18) [19] (p. 323) [7] (p. 292f). Furthermore, it is important to observe and accurately measure the dependent variable (potential effect) [25] (p. 7) [26] (p. 43). Many students, however, have been shown to have misconceptions. For example, students may think that the goal of an experiment is to create an effect (engineering mode) instead of examining causal relationships [21] (p. 12) [27] (p. 860ff.). The experiment has to be precisely described so that it can be repeated [25] (p. 7). It requires consideration of appropriate methods and conducting experiments in a standardized way [25] (p. 7).

Finally, the data must be analyzed precisely. When data are assessed, errors have to be analyzed and taken into account [28] (p. 155). Furthermore, students should "differentiate experimental error [. . .] from experimental effect" [21] (p. 7). Results are compared with the hypothesis, which is accepted, rejected, or further examined [21] (p. 9) [19] (p. 324). It is essential to "guard against one's own confirmation bias in data interpretation" [21] (p. 7). Confirmation bias can influence the reasoning in that specific data that do not support the hypothesis are ignored. It can be difficult for students to reject a hypothesis due to their beliefs [21] (p. 9) [24] (p. 84f.).

Learning outcomes relevant to experimentation are prescribed in the German National Educational Standards [4]. According to the standards, students are expected to be able to plan, conduct, and analyze experiments at the end of grade 10 [4] (p. 14). Teachers have to know the learning goals and understand the student conceptions in order to conduct

assessments that serve learning [2] (p. 2, 10). Therefore, this situation requires CK and PCK, i.e., knowledge of experimentation and student conceptions in biology.

Constructs that are related to the *knowledge of what to assess* regarding experimentation competences are *examining competence* and *diagnostic competence*. *Examining competence*, similar to experimentation competences, comprises the facets of questions, hypotheses, design and performance, and analysis and interpretation, which has been the focus of analysis for pre-service science teachers [29] (p. 40ff.). Assessments of experimentation competences that serve to learn have only been focused on by a few research studies so far. One example is a study conducted with biology pre-service teachers: Dübbelde [14] investigated the *diagnostic competence* for experimentation competences. The tasks performed in a closed answer format captured pre-service teachers' ability to apply given criteria, such as linking the conclusion to the hypothesis, but not the *knowledge of what to assess*.

Besides *knowledge of what to assess*, efficacy beliefs can influence the performance in our test. High self-efficacy beliefs can enhance the useful application of knowledge [30,31] (p. 211). Personal teaching efficacy of student interns correlated, e.g., with their lesson presenting behavior and questioning behavior [32] (p. 413).

Studies of *scientific reasoning*, *scientific inquiry*, and *experimentation competences* described conflicting findings regarding the dimensionality of the constructs [22,33,34]. Weak and intermediate latent correlations of 0.33–0.73 between the subscales related to question, hypothesis, planning, and interpretation (condensed label of subscales used by the authors) indicate that different skills are necessary for the different phases of *scientific reasoning* [33] (p. 58). Wellnitz's study of *scientific inquiry*, on the contrary, found higher latent correlations between the scales question, hypothesis, experimental design, and data analysis (0.80–0.95) [22] (p. 132) so that the authors of this study argue that comprehensive skills are necessary for all phases of experimentation.

Teachers need to possess experimentation competences to be able to evaluate student achievement. In particular, explicit knowledge of criteria and misconceptions enables teachers to assess student achievement against curricular expectations. When experimenting in class, teachers can focus on one of the three phases of experimentation. To convey an understanding of scientific inquiry, however, it is helpful for students to engage themselves in the whole process [25] (p. 5). Hence, teachers should have an understanding of all three phases of experimentation.

6.1.3 Research Questions and Hypotheses

The goal of the study is to model and measure pre-service biology teachers' knowledge and skills regarding the assessment of high school students' experimentation competences. Depending on theoretical background, two different models of *knowledge of what to assess* regarding experimentation competences in biology can be derived: a one-dimensional (1D) model comprising the three phases of experimentation, and a three-dimensional (3D) model taking into account the different requirements for forming hypotheses, planning experiments and analyzing data. By modeling and measuring it can be learned more about the dimensionality and quality of pre-service biology teachers' assessment literacy on *what to assess*. Therefore, a reliable and valid measurement instrument is necessary.

This type of query led to three research questions:

The first question concerns the construct dimensionality and test quality.

1. In what way can *knowledge of what to assess* regarding experimentation competences in biology be modeled and measured?

For investigating the validity of our conclusions, the following constructs related to *knowledge of what to assess* regarding experimentation competences are relevant: Given the knowledge and skills that are necessary to assess experimentation competences, *examining competence* and *diagnostic competence* for experimentation competences should be closely related to the construct measured in our study. Moreover, an analysis of correlations between *knowledge of what to assess* and learning outcomes as well as an analysis of differences between known groups is interesting regarding validation, leading to the second research question.

2. To what extent is the *knowledge of what to assess* related to similar constructs and learning outcomes? To what extent can differences be found in the *knowledge of what to assess* between students at the undergraduate and graduate levels?

Regarding research question two, we expect correlations between *knowledge of what to assess* and *diagnostic competence* as well as *examining competence.* Both, the instrument for *diagnostic competence* regarding experimentation competences and the instrument for *examining competence* share a focus on experimentation with our construct. A lower correlation than between *knowledge of what to assess* and *diagnostic competence* and *examining competence* is expected between *knowledge of what to assess* and *self-*

efficacy beliefs regarding teaching biology since *self-efficacy beliefs* are based on a broader range of knowledge than the *knowledge of what to assess* regarding experimentation competences.

We expect correlations between *knowledge of what to assess* regarding experimentation competences and grades as an indicator for learning outcomes in high school biology, in biology at university, and biology teacher education courses at university. Since biology teacher education courses can deal with assessment and student conceptions, the grade in biology teacher education is expected to correlate highest with our construct. Moreover, we expect correlations between *knowledge of what to assess* regarding experimentation competences and the number of respective learning opportunities.

Students at the graduate level are hypothesized to outperform students at the undergraduate level because the former are expected to have acquired more *knowledge of what to assess* during their teacher education studies than students at the undergraduate level. Assessment, knowledge and skills in experimentation, and knowledge of student conceptions in biology are prescribed contents for biology teacher education [12]. Therefore, we hypothesize the following:

Students at the graduate level reach higher person abilities in the *knowledge of what to assess* regarding experimentation competences than students at the undergraduate level.

Once a reliable and valid measurement instrument has been developed, the third research question aims at providing information about pre-service teachers' *knowledge of what to assess*.

3. What are the strengths and weaknesses of pre-service biology teachers regarding *knowledge of what to assess* regarding experimentation competences in biology?

6.2 Methods

6.2.1 Participants and Data Collection

The study was conducted from October 2014 to February 2015, including pre-service biology teachers from 18 German universities in seven federal states. We analyzed questionnaire answers of $n = 495$ pre-service biology teachers (78.1% female, mean age = 23.15 years, $SD = 3.20$ years; the gender distribution represents the higher percentage of female pre-service teachers in Germany). Five people of $N = 500$ were excluded from analyses due to missing data or improper handling of the questionnaire. The participants of the

study covered a range of different semesters in Bachelor, Master, or State Examination studies (34.3% Bachelor, 41% Master, 24.7% State Examination). In the following the term *students at the undergraduate level* comprises students in their Bachelor studies as well as students striving for the State Examination degree ≤ semester 6. The term *students at the graduate level* comprises students in their Master studies as well as students striving for the State Examination degree ≥ semester 7. The study participants were seeking to become primary, secondary, and vocational school teachers or special education teachers. In Germany, both the Master and First State Examination degree qualify for teaching practice in the second phase of teacher education.

Data were collected using a paper-pencil questionnaire which recorded (a) demographic and academic information, (b) *knowledge of what to assess* regarding experimentation competences in biology, and (c) *diagnostic competence* or *self-efficacy beliefs* for teaching biology. An instrument to measure *examining competence* [35] was part of a parallel conducted study focusing on teaching competences for experimentation in biology [36]. The study on teaching competences for experimentation in biology and our study has an overlapping sample of pre-service teachers who answered both questionnaires. Two research associates and one student assistant surveyed data collection using a standardized procedure at 18 universities in seven federal states.

6.2.2 Measurement Instrument

For the measure of *knowledge of what to assess* regarding experimentation competences in biology, we built on the instrument of Bögeholz et al. [36], keeping well-functioning items and shortening the instrument to not exceed 90 minutes in testing time. These measures facilitated testing in sessions of seminars and made the testing time acceptable for pre-service teachers outside of seminars. Furthermore, it supported (test) performance by preventing a decrease in motivation and an increase in fatigue [37]. Thus, seven out of the initial 27 scenarios portraying different phases and competences of experimentation were chosen and adapted accordingly from Bögeholz et al. [36].

Each of the seven scenarios described an experimentation assignment for a biology lesson with hypothetical high school students and the response of a single student or a group of students (Figure 6-1). Pre-service teachers were asked to assess the response of the hypothetical student(s). For some scenarios, they had to explain the student conception that influenced his/her procedure and in some cases to correct the solution in addition. The applied contexts covered the required basic curricular content. Relevant information

for the experiments was given so that no additional content knowledge about the contexts was required.

The measurement instrument for *knowledge of what to assess* regarding experimentation competences in biology consisted of seven biology lesson scenarios (see Figure 6-1 for an example) covering the phases of hypothesis formation, design of an experiment, and analysis of data. Each phase was focused on at least in two scenarios in different contexts that were chosen in consideration of German core curricula for biology [38].

Mrs. Nell discusses the characteristics of enzymes with her students in class 10. She asks her students to examine at which temperature α-amylase breaks down starch the fastest.

Mrs. Nell gives instructions on how to design the experiment:
Three test tubes are each filled with the same amount of starch solution. Then the starch solutions in the test tubes are placed in different water baths to reach the desired temperatures: 10°C, 40°C, and 70°C. To each starch solution, α-amylase of the same temperature (10°C, 40°C, 70°C) is added. α-amylase breaks down starch into maltose and glucose. Then every minute, a drop of the starch solution of every test tube is taken and added to brown iodine solution. When starch is added to the iodine solution, the color of the mixture changes from brown to blue. When maltose and glucose are added to the iodine solution, the brown color does not change.

Steps conducted by the student Bea:

Bea's hypothesis

α-amylase breaks down starch the faster, the higher the temperature of the starch solution is.

Bea's design and performance

Test tube 1: Starch solution, α-amylase, 10°C
Test tube 2: Starch solution, α-amylase, 40°C
Test tube 3: Starch solution, α-amylase, 70°C

A drop of each solution is added after 1, 2, 3, 4, and 5 minutes to 2 drops of iodine solution, respectively.

Results of the experiment

	1 min	2 min	3 min	4 min	5 min
10°C	*blue*	*blue*	*blue*	*brown*	*brown*
40°C	*blue*	*brown*	*brown*	*brown*	*brown*
70°C	*blue*	*blue*	*blue*	*blue*	*blue*

blue: Starch is still present
brown: Starch has been broken down into maltose and glucose

Bea's conclusion

The activity of α-amylase increases with rising temperature.

Tasks for pre-service teachers:

1. Assess Bea's data analysis. Give reasons. (Item 15)
2. Explain how Bea could have come to her conclusion. (Item 16)

Figure 6-1 Biology lesson scenario with assessment tasks for pre-service teachers (slightly adapted layout).

The context seed germination (scheduled for grades five and six) was represented in three scenarios. The contexts photosynthesis (scheduled for grade seven and eight) and enzymology (scheduled for grade nine and ten) were each represented in two scenarios [38]. The composition of the questionnaire is shown in the matrix of Table 6-1. The corresponding item list is displayed in Table 6-10 in Appendix.

Table 6-1 Matrix of contexts x phases of experimentation with scenarios and corresponding items (see Table 6-10 in Appendix).

	Seed Germination	Photosynthesis	Enzymology
Hypothesis formation	Scenario 1 with items 1, 2, 3	Scenario 4 with items 4, 5, 17, 18	
Design of an experiment	Scenario 2 with items 6, 7, 19	Scenario 5 with items 13, 14, 20	Scenario 6 with items 8, 9, 10
Analysis of data	Scenario 3 with items 11, 12		Scenario 7 with items 15, 16

The task format required study participants to assess students' experimentation competences according to central criteria. The following criteria were used for the phase of hypothesis formation: comprehensive hypothesis formation and, concerning single hypotheses, being testable and founded. Regarding the phase of designing an experiment, the following criteria were used: systematic variation of variables and precise design. Furthermore, the following criteria were used for assessing the planning of the performance: accurate measurement procedures and standardization. Concerning the phase of data analysis, the following criteria were used: correct data analysis, precise data analysis, error analysis, and conclusion with a link to the hypothesis.

Moreover, the two student conceptions engineering mode of experimentation and confirmation bias had to be assessed. The implementation of the criteria of all three phases and student conceptions was realized in different categories of items: assessing student conceptions, assessing correct student solutions, and assessing incorrect student solutions. Because the tasks measuring pre-service biology teachers' assessment literacy were based on scenarios, we expected them to have curricular validity and to be motivating. The task format was close to real-world performance tasks and focused on the criteria for experimentation that are relevant for learning outcomes in biology at high school.

6.2.3 Coding of Knowledge of What to Assess Regarding Experimentation Competences

For each biology lesson scenario, two to four items were coded (see Table 6-2 and Table 6-3 and Table 6-10 in Appendix). The coding was a further development of the coding applied in the pilot study [36]. It was equally distributed to four persons and carried out according to a manual which was deductively and inductively developed [39]. The scoring of the answers to the tasks considered correctness, completeness, and accuracy (Table 6-10). Ten trichotomous items had a maximum score of 2 (scores 0, 1, 2) (Table 6-2 and Table 6-3); ten dichotomous items had a maximum score of 1 (0, 1). For the dichotomous items, the maximum score was relativized to 2, assigning all items the same weight. A randomly chosen representative tenth of the test booklets, i.e., 52 test booklets, was analyzed by all four persons to investigate the inter-coder reliability. A sufficient power of kappa was reached with this sub-sample [40]. However, an analysis of Krippendorff's alpha was preferred for ordinal data. Krippendorff's alpha was analyzed for the most differentiated version of the scoring rubrics before item steps were combined. Four of the 20 items reached a Krippendorff's alpha below 0.70. For the other 16 items, Krippendorff's alpha was between 0.70 and 0.87. A low Krippendorff's alpha could be explained by the open-ended tasks and the original superfine scoring. After combining item steps, it can be assumed that Krippendorff's alpha improved [41].

Table 6-2 Scoring of Item 15 (experimentation phase: analysis of data, criterion: incorrect data analysis) – task of item 15: "Assess Bea's data analysis. Give reasons." (cf. Table 6-10 in Appendix).

	Scoring	Exemplary answers
Score 2	The criterion is named and explained.	The data analysis is wrong. No transformation could also be detected at 70°C. (1.11) Bea's data analysis is not detailed enough. The efficiency of α-amylase increases up to 40°C, but above that, no splitting takes place at all. Therefore, Bea's conclusion is wrong. (1.13)
Score 1	The criterion is named.	The data analysis is incomplete since not all data have been taken into account. (1.9) Her conclusion is wrong. It is possible that the relationship is not clear to her: that higher enzyme activity can explain the splitting of starch and therewith the change to a brown color. (1.17)
Score 0	The criterion is neither named nor explained.	The table would have been better the other way around. 10°C 40°C 70°C 10°C 40°C 70°C 1 min x x x 4 min o 2 min x 5 min o 3 min x (1.1) The data analysis in a table is good. The intervals increase constantly and everywhere equally. (1.7)

Table 6-3 Scoring of Item 16 (experimentation phase: analysis of data, criterion: confirmation bias) – task of item 16: "Explain how Bea could have come to her conclusion." (cf. Table 6-10 in Appendix).

	Scoring	Exemplary answers
Score 2	The criterion is explained completely. The explanation includes both of the following aspects: (1) student ignores the observation (of the 70°C test tube) OR the student does not consider the result (of the 70°C test tube) due to certain reasons. (2) student has a specific belief concerning the outcome of the experiment OR the student tends to confirm the hypothesis.	Bea looks for clues that confirm her hypothesis. She ignores other results of her experiment since they don't fit her belief. (confirmation bias effect?) (1.16) She might conclude, due to previous knowledge, that reactions take place faster at higher temperatures. With the experiment, she verifies her own expectations and ignores contradicting results. (1.70)
Score 1	The criterion is explained in parts. The explanation includes one of the two following aspects: (1) student ignores the observation (of the 70°C test tube) OR the student does not consider the result (of the 70°C test tube) due to certain reasons. (2) student has a particular belief concerning the outcome of the experiment OR the student tends to confirm the hypothesis.	Bea ignored the results of the 70°C test tube. (1.13) Bea might have only compared the 10°C and 40°C and excluded 70°C as a mistake. (1.111)
Score 0	The criterion is not explained.	Maybe she read her table falsely. To the right there are more and more brown fields that indicate that starch has been broken down. (1.1) Bea might have mixed up the variables time and temperature in her statement. (1.9)

6.2.4 Validation Instruments

In addition to demographic and academic information, *diagnostic competence*, *examining competence* and *self-efficacy beliefs* for teaching biology, were measured for validation purposes, each for a sub-sample.

Diagnostic competence for experimentation competences in biology was assessed with an instrument developed by Dübbelde [14]. This instrument was shortened from 17 to 12 items for the use in our study. The original 17 and remaining 12 items dealt with central conditions for experimentation, such as the foundation of the hypothesis, distinction between observations and conclusions, or link of conclusion to the hypothesis. The instrument consisted of hypothetical educational materials and products, i.e., high school students' worksheets and students' notes taken during an experiment, and an assessment

sheet for pre-service teachers with 12 items focusing on the phases hypothesis formation, design of an experiment, performance of the experiment in the sense of documentation and analysis of data. In the items, the pre-service teachers had to indicate whether certain conditions of experiments, such as performance of error analysis, had been fulfilled by the hypothetical students (nine items: "yes", "no", and "don't know") or identify the correct answer out of four choices (one item), out of three choices (one item) or out of three options, among that "don't know" (one item). The closed answer format (two choices plus "don't know") led to a high probability of guessing. The instrument with the original 17 items reached a Cronbach's alpha of 0.50 [14] (p. 189). The shortened instrument with 12 items that we applied for validation purposes had a Cronbach's alpha of 0.36 (n = 136). The instrument on *diagnostic competence* shared the focus on the assessment of students' experimentation competences with our instrument on *knowledge of what to assess*. However, the instrument of Dübbelde [14] asked for the estimation of the given criteria allowing guessing. The instrument did not focus on the personal *knowledge* of pre-service teachers on *what to assess*. Besides giving the criteria for experimentation, the *diagnostic competence* instrument differed from ours in that the instrument tested neither the knowledge of student conceptions nor the correction of specific incorrect hypothetical student solutions.

Examining competence in biology was assessed using a short scale (12 multiple-choice items) developed by Krüger et al. [35]. The instrument included the experimental phases of question formation, hypothesis formation, design of experiments, and analysis of data. Pre-service teachers had to select either a suitable question for an examination, a hypothesis that can be derived from observation, a hypothesis that is the basis of the examination, a design for the experiment that is suitable to test a specific hypothesis, or the correct data analysis of the experiment. For all choices to be taken, one out of four answers was correct. Criteria for experimentation, such as holding the independent variables constant in an experiment, have to be applied to select the correct answer. Moreover, the instrument captures contents of the knowledge base for the assessment of experimentation competences. It differs from our test on *knowledge of what to assess* regarding experimentation competences in that the criteria for experimentation do not have to be named, explained, or described. A "feeling" for how to design an experiment, for instance, is sufficient to solve the tasks. And again, guessing can also lead to the correct answer, up to 25% of the time. The instrument reached a Cronbach's alpha of 0.39 (n = 239) in our study.

The third instrument applied for validation purposes measured pre-service teachers' *self-efficacy beliefs* for teaching biology. On a Likert scale, pre-service teachers had to indicate their expected abilities concerning, for instance, planning and conducting lessons in consideration of research results on biology education, such as research results regarding student conceptions (four items) and planning lessons in consideration of core concepts ("Basiskonzepte") of biology, such as structure and function, and competences for biology (two items) [42]. Both, research results on biology education, as well as core concepts and competences for biology, comprise information relevant for experimentation in the classroom: The ability to plan and conduct lessons in consideration of research results on biology education includes the knowledge of and ability to use research findings on students' biological conceptions. The competences for biology comprise experimentation competences.

6.2.5 IRT Modeling and Further Analyses

Data analysis was conducted using the partial credit model [43]. Item Response Theory (IRT) analyses were conducted with ConQuest [44]. For item related analyses, the average person's ability was set to zero (=case-centered analysis, constraints = cases). Due to this procedure, also the item difficulty of the last test item could be estimated correctly. For person related analyses, the average item difficulty was set to zero (=item-centered analysis, constraints = items) [44].

The data quality was checked for the one- and 3D model via fit statistics ($0.8 \leq$ wMNSQ ≤ 1.2; $-2 \leq$ t-value ≤ 2) resulting from case-centered IRT analyses [45] (p. 164 ff.) [46] (p. 270ff.). Item-centered analyses were conducted to estimate person-measures and compare the fit of the two models. The deviance, as well as Bayesian information criterion (BIC) and Akaike's information criterion (AIC) and latent correlations between the dimensions, were computed. An analysis of differential item functioning (DIF) was conducted with ConQuest to identify items that were biased for the educational level or gender.

For validation, *knowledge of what to assess* regarding experimentation competences and the related constructs *diagnostic competence* with 12 items [14] and *examining competence* with 12 items [35] were analyzed by multidimensional modeling, and latent correlations were examined. The multi-dimensional case-centered analysis (one dimension for each of the three constructs above) provided fit statistics for the items of the three scales. Moreover, manifest correlations between *knowledge of what to assess* regarding experimentation competences and different *self-efficacy beliefs* concerning *planning and conducting lessons in consideration of research results on biology education* and *planning*

lessons in consideration of core concepts and competences for biology [42] were analyzed. Correlations between *knowledge of what to assess* regarding experimentation competences, grades, and the number of learning opportunities, were computed for a further check of validity. In addition, a Mann-Whitney-U-test was applied with person measures to examine whether students at the graduate level outperform students at the undergraduate level in their *knowledge of what to assess* regarding experimentation competences in biology. Using item difficulties, we examined which criteria of experimentation are easy or difficult to assess for pre-service biology teachers (strengths and weaknesses of pre-service teachers). We compared item difficulties of different item groups using one-way ANOVA and a post hoc Tukey HSD test for specific group comparisons. In sum, we used the steps of Figure 6-2 to analyze the data.

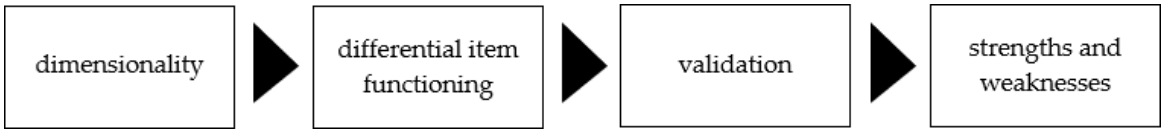


Figure 6-2 Foci of data analyses of pre-service teachers' knowledge of what to assess.

6.3 Results

6.3.1 Modeling and Measuring Knowledge of What to Assess Regarding Experimentation Competences

Dimensionality

The case-centered 1D modeling of *knowledge of what to assess* regarding experimentation competences as well as 3D modeling with the dimensions hypothesis formation, design of an experiment, and analysis of data reached comparable reliabilities and item parameters. The EAP/PV for the 1D model was 0.60 and lay between 0.50 and 0.54 for the three dimensions of the 3D model (hypothesis formation: 0.50, design of experiments: 0.54, analysis of data: 0.51). The item fit was good for both models: The wMNSQ values ranged from 0.92 to 1.06 in the 1D modeling, and the corresponding t-values ranged from −0.9 to 1.7. The 3D modeling yielded wMNSQ values of 0.92–1.08 and t-values of −0.8–2.0.

Comparing the fit of the 1D and 3D model, item-centered analyses revealed the following (Table 6-4): the BIC that considers the model complexity indicated a better fit for the 1D model. Regarding the deviance and AIC, the 3D model (deviance = 13,236, *AIC* = 13,335.08) fit (slightly) better to the data than the 1D model (deviance = 13,279, *AIC* =

13,340.76). The latent correlations between the three dimensions ranged from 0.57 to 0.80 (Table 6-5). These rather low latent correlations indicated that the three dimensions captured different knowledge dimensions. Considering the construct that should be measured, however, the 1D modeling was more appropriate than 3D modeling. It covered *knowledge of what to assess* regarding experimentation competences more comprehensively and with an adequate number of items and was therefore applied for the following analyses.

Table 6-4 Comparison of the 1D and 3D model (item-centered analysis, n = 495).

Models	Deviance	Parameter	BIC	AIC
1D	13,279	31	13,471.10	13,340.76
3D	13,263	36	13,486.45	13,335.08

Table 6-5 Item-centered analysis of latent correlations between hypothesis formation, design, and analysis of data of knowledge of what to assess regarding experimentation competences (n = 495).

	Hypothesis Formation	Design
Hypothesis formation	---	
Design	0.68	---
Analysis of data	0.57	0.80

Test and Item Parameters

The case-centered 1D IRT modeling revealed acceptable reliabilities (EAP/PV reliability = 0.60, item separation reliability = 0.994) (Table 6-6). The variance of 0.14 indicated that the differentiation between persons was low. As stated in the section "dimensionality", the item fit was good (0.8 ≤ wMNSQ ≤ 1.2; −2 ≤ t-value ≤ 2). The item difficulties of the 1D model ranged from −1.34–2.22 logits. All item steps had been reached by at least 5% of the pre-service teachers, except for Item 6, step 2. The item step was maintained due to the relevance of the content: knowledge of the student conception engineering mode of experimentation. The discrimination of the items reached acceptable values above 0.25 [47] (p. 147) except for Item 6, focusing on the student conception engineering mode of experimentation (0.15) and Item 18 dealing with the correction of an unfounded hypothesis (0.21). Both items were kept due to the relevance of their content.

Table 6-6 Parameters of the case-centered 1D IRT modeling of knowledge of what to assess regarding experimentation competences.

	1D model
Total Number of items (dichotomous/trichotomous)	20 (10/10)
EAP/PV reliability, item separation reliability	0.60, 0.99
Variance	0.14
Item difficulty: min to max	−1.34-2.22
Person ability: min to max	−2.85-1,33
wMNSQ: min to max	0.92-1.06
T value: min to max	−0.9-1.7
Discrimination: min to max	0.15-0.45

Differential Item Functioning

Considering the 1D modeling of *knowledge of what to assess* (scale with 20 items), students at the undergraduate level (n = 253) scored 0.32 logits lower than students at the graduate level (n = 224). Differential item functioning (DIF) existed for Item 7 unsystematic variation of variables (logit difference = 0.74). Item 7 was the easiest item for students of both groups. It was considerably easier for students at the graduate level (solved by 98% of students at the graduate level and 87% of students at the undergraduate level). The logit difference for all other items was below 0.4. Thus, it was not regarded as a considerable DIF [48] (p. 12). No considerable DIF occurred for gender (maximum logit difference = 0.28).

The Wright Map (Figure 6-3) shows that nine items/item steps out of 30 items/item steps were complicated. Consequently, they did not differentiate very well between person abilities. For the range of −1.00–0.70 logits, the distribution of the item difficulties matched the person's abilities well, except for a minor gap of items between Item 8 and 9.

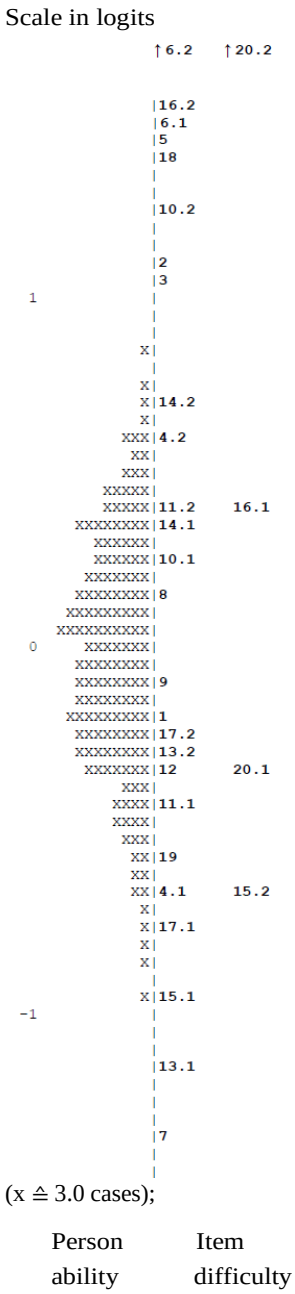


Figure 6-3 Wright Map of the case-centered 1D IRT modeling of knowledge of what to assess regarding experimentation competences (↑ = the item difficulty is greater than presented).

6.3.2 Validation of Knowledge of What to Assess with Related Constructs, Educational Outcomes, and Comparison of Known Groups

Relationship to Related Constructs

The 3D (case-centered) IRT modeling of *knowledge of what to assess* regarding experimentation competences, *diagnostic competence*, and *examining competence* showed

that the fit of all items of the three instruments was good (0.8 ≤ wMNSQ ≤ 1.2; −2 ≤ t-value ≤ 2).

Analyses revealed that pre-service teachers reached the highest person measures for *diagnostic competence* (mean person ability = 1.10) (Table 6-7). Lower person measures were reached for *examining competence* (mean person ability = 0.16) and the lowest for *knowledge of what to assess* regarding experimentation competences (mean person ability = −0.14). Thus, it was the most difficult construct. Table 6-8 shows the latent correlations between the constructs.

Table 6-7 Parameters of the item-centered 3D IRT modeling (n = 128).

	Mean	Variance	EAP/PV
Knowledge of what to assess	-0.14	0.09	0.58
Diagnostic competence	1.10	0.27	0.45
Examining competence	0.16	0.29	0.59

Table 6-8 Latent correlations between the three constructs (item-centered analysis) (n = 128).

	Knowledge of What to Assess
Diagnostic competence	0.37
Examining competence	0.78

The highest latent correlation existed *between knowledge of what to assess* regarding experimentation competences and *examining competence* (0.78). The latent correlation between *knowledge of what to assess* regarding experimentation competences and *diagnostic competence* was relatively low (0.37).

The analysis of correlations (Spearman) between *knowledge of what to assess* regarding experimentation competences and *self-efficacy beliefs* for teaching biology revealed the following results: *Knowledge of what to assess* regarding experimentation competences of students at the graduate level correlated with their *self-efficacy beliefs* regarding the ability to *plan and conduct lessons in consideration of research results on biology education* s ($r = 0.20$, $p < 0.05$, $n = 146$) as well as *self-efficacy beliefs* regarding the ability to *plan lessons in consideration of core concepts and competences for biology* ($r = 0.22$, $p < 0.01$, $n = 147$). In contrast, no relationship between the variables was found for students at the undergraduate level ($p > 0.05$; $n = 178$, $n = 181$).

Relationship to Grades and Learning Opportunities

Table 6-9 shows the correlations of person abilities in the *knowledge of what to assess* regarding experimentation competences with educational variables. Better grades in high school biology as well as in courses of biology and biology teacher education at university correlated positively with person measures: The more achieved points at high school (r = 0.19, $p < 0.01$) and the lower (that is, the better) the university grade in biology ($r = -0.16$, $p < 0.01$) and biology teacher education ($r = -0.28$, $p < 0.01$), the higher were the person abilities. There was a strong correlation between university grades in biology and biology teacher education ($r = 0.63$, $p < 0.01$).

The amount of learning opportunities correlated with *knowledge of what to assess* regarding experimentation competences: The more courses in biology teacher education pre-service teachers had completed, the higher the person abilities in the *knowledge of what to assess* regarding experimentation competences.

Table 6-9 Correlations between knowledge of what to assess and educational variables.

	High School	**University**		
Variable	Last grade in biology in high school	Average grade in university courses in biology	Average grade in university courses in biology teacher education	Number of completed courses in biology teacher education
Person ability	0.19²*s* (n = 446)	-0.16²*s* (n = 377)	-0.28²*s* (n = 265)	0.21²*p* (n = 406)

Legend. s = Spearman, p = Pearson, ² = $p < 0.01$; person ability: test result (20 items of *knowledge of what to assess*); last grade in biology in high school: 1 = very poor, up to 15 = very good; average grade in courses in biology as well as biology teacher education: 1.0–1.3 = very good, 1.7–2.3 = good, 2.7–3.3 = satisfactory, 3.7–4.0 = sufficient; the number of completed courses in biology teacher education: 1 = 1, up to 10 = 10.

Comparison of Known Groups

Regarding the 1D model of *knowledge of what to assess* regarding experimentation competences, Q-Q-plots indicated that data were normally distributed for the students at the graduate level ($n = 224$) but not for students at the undergraduate level ($n = 254$, negatively skewed). The Levene-test indicated that homogeneity of variances could not be assumed for the two groups ($p = 0.025$). The Mann-Whitney U-test was applied: There was a statistically significant difference in person abilities between students at the undergraduate level (M_{Rank} = 207.25) and at the graduate level (M_{Rank} = 276.07), U = 20,255.50, Z = −5.447, $p < 0.001$, with a moderate effect size ($r = 0.25$). Consequently, the person ability increased in the course of academic studies (Figure 6-4), which is in line with our hypothesis.

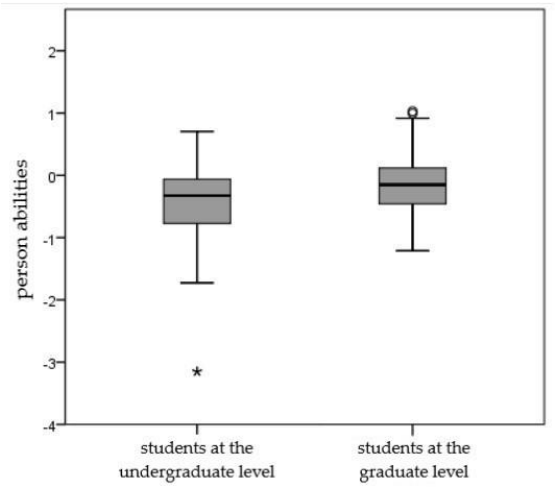


Figure 6-4 Person abilities of students at the undergraduate and graduate level (item-centered analysis, 1D modeling; undergraduate level: Bachelor + State Examination degree ≤ semester 6, graduate level: Master + State Examination degree ≥ semester 7).

6.3.3 Strengths and Weaknesses Concerning Knowledge of What to Assess Regarding Experimentation Competences

Analyzing the distribution of items on the Wright Map (Figure 6-5), we were able to identify specific contents that influence item difficulty. We were able to group these contents into four categories. Category i focusses on the assessment of student conceptions. Category ii deals with the assessment of correct student solutions. The further two categories comprise the assessment of incorrect student solutions: Category iii focusses on the assessment of the planning of the performance with regard to standardization and accuracy and

category iv on the assessment of further incorrect student solutions. In the following, we describe the contents of the four categories in order of decreasing difficulty.

Ad (i) Student conceptions were displayed in the trichotomous Item 6 (2.22 logits, Figure 6-5) and Item 16 (0.98 logits). Item 6 focused on the student conception engineering mode of experimentation. Item 16 dealt with the student conception confirmation bias. The assessment of student conceptions was very difficult. Only a few pre-service teachers named and explained the engineering mode of experimentation (full credit) or explained the confirmation bias comprehensively (full credit).

Ad (ii) The assessment of correct student solutions included the assessment of a testable hypothesis (Item 2), a theoretically founded hypothesis (Item 3), and the systematic variation of variables (Item 8). The item difficulties of this group of items ranged from 0.15 to 1.10 logits.

Ad (iii) The assessment of criteria concerning the planning of a standardized and accurate procedure was difficult but less complicated than assessing student conceptions and correct student solutions. The item difficulty for Item 10, which required naming that several aspects lacked standardization, was 0.77 logits. The item difficulty for Item 14, which required naming and explaining that the measurement procedure was inaccurate, was 0.53 logits.

Ad (iv) In comparison, the assessment and correction of incorrect student solutions were relatively easy: The item difficulty of Item 4 dealing with the assessment of an untestable hypothesis was 1.11 logits below the item difficulty of Item 2 requiring the assessment of a testable hypothesis. The correction of an untestable hypothesis (Item 17) was even easier. The assessment (Item 7) and correction (Item 19) of an unsystematic variation of variables was considerably easier than the assessment of the systematic variation of variables (Item 8). The logit difference in Items 7 and 8 was 1.49 (Figure 6-5). Further items required the assessment of incorrect student solutions, i.e., incomprehensive hypothesis formation (Item 1, −0.15 logits), imprecise design of experiment (Item 9, −0.09 logits), imprecise data analysis (Item 11, −0.02 logits), missing error analysis (Item 12, −0.30 logits), conclusion without a link to hypothesis (Item 13, −0.72 logits) and incorrect data analysis (Item 15, −0.80 logits) were relatively easy. The whole item group had item difficulties ranging from −1.34 to −0.02 logits.

There were three exceptions regarding this item category: Item 5 and 18 dealing with the assessment and correction of an unfounded hypothesis, i.e., a hypothesis without justification (Item 5, 1.43 logits; Item 18, 1.40 logits) and Item 20 dealing with the correction of

a conclusion without a link to the hypothesis (1.01 logits). The difficulties of Item 5 and 18 could have been influenced by the task format that required assessing (Item 5) or correcting (Item 18) several mistakes for one task, namely the missing foundation of the hypothesis (Item 5 and 18) in addition to the missing testability of the hypothesis (Item 4 and 17). Item 20 required the formulation of a conclusion with reference to the hypothesis and its verification for full credit. Few pre-service teachers went beyond the correction of the content of the given hypothetical student answer (scored with partial credit) and verified the hypothesis. No statement regarding the missing verification of the hypothesis was required for full credit of the 1.73 logits easier Item 13 covering the assessment of a conclusion without a link to the hypothesis. These three items were exceptional cases as they required demanding information processing for generating additional solutions. The task format could have influenced the item difficulty. Therefore, we excluded them from the item category iv.

A one-way ANOVA revealed that the four item categories differed significantly ($F(3, 13) = 13.64$, $p < 0.001$). A post hoc Tukey HSD test indicated that the mean item difficulty for the items of category iv (item group iv) assessing incorrect student solutions (mean = −0.45, $SD = 0.42$) differed significantly from the three other item groups (vi versus i: mean = 1.60, $SD = 0.88$, $p = 0.001$; vi versus ii: mean = 0.77, $SD = 0.54$, $p = 0.009$; vi versus iii: mean = 0.65, $SD = 0.17$, $p = 0.048$). The other three item groups i, ii, and iii did not differ significantly in their mean item difficulty from each other ($p > 0.05$).

Regarding the three phases of experimentation, no significant differences between mean item difficulties of the three phases were found, considering all 20 items ($p > 0.05$; Figure 6-5).

Overall, relatively few pre-service teachers used certain scientific terms, such as engineering mode, confirmation bias, error analysis, and control group, to assess the hypothetical high school students' experimentation competences. Most of them described the criterion without naming these specific terms. For instance, regarding item group i, 76 pre-service teachers identified the idea of the underlying engineering mode of experimentation (evaluated by partial credit), only seven of them used the scientific term engineering mode. Forty-nine pre-service teachers assessed the confirmation bias; only one of them named the student misconception confirmation bias. Regarding item group iv, the idea of missing error analysis was perceived by 313 pre-service teachers and the term named by only 34 of them. In the assessment of the unsystematic variation of variables, only 94 of 456 pre-service teachers used the term control group.

Figure 6-5 Wright Map of the case-centered 1D IRT modeling of knowledge of what to assess regarding experimentation competences (without item steps) (↑ = the item difficulty is greater than presented).

6.4 Discussion

In the following section, the results are discussed in the order of the research questions focusing on dimensionality and test quality, validation, and strengths and weaknesses of pre-service biology teachers.

6.4.1 Dimensionality and Test Quality

Concerning dimensionality, one could argue for a 1D model or a 3D model considering the modeling with the partial credit model. Taking into account the construct *knowledge of what to assess* regarding experimentation competences, a 1D model is the preferred option because of considerations set out in the following: The 1D parsimonious approach regarding the competence construct in science education represented an advantage for empirical testing cf. [49] and for teaching and assessing taking into account the other competences to be learned by pre-service teachers. For instance, pre-service teachers also have to learn to analyze and plan lessons to foster high school students' experimentation competences [36] and in the frame of assessment literacy they have to acquire knowledge of assessment purposes, knowledge of assessment strategies, and knowledge of assessment interpretation and action-taking [16]. Thus, we prioritized a more manageable conceptualization as opposed to more differentiated analyses possible with a more complex competence construct for *knowledge of what to assess* cf. [49] (p. 63). The benefits of operating with a broader construct for practical usefulness outweighed, in this case, a more differentiated conceptualization. Thus, arguments for multi-dimensionality such as the given latent correlations receded into the background.

The phenomenon that empirical results regarding experimentation related knowledge did not provide a clear picture concerning dimensionality was not only given for the construct *knowledge of what to assess* in the group of pre-service teachers. For example, varying results regarding dimensionality occurred in the research on similar constructs such as *scientific inquiry*, *scientific reasoning*, and *experimentation competences* investigating high school students. We summarize this research and structure the summary by proceeding from the more advanced students to the less advanced students: (i) Research with 10th graders on *scientific inquiry* revealed high latent correlations between the scales question, hypothesis, design, and data analysis (0.80–0.95) [22] (p. 132). It turned out that neither a four-dimensional model (comprising question, hypothesis, design, and data analysis) nor a 3D model (comprising observing, comparing, experimenting) outweighed a 1D model of *scientific inquiry*, which was in line with the approach to operate with a manageable amount of competence models for teaching biology. (ii) An analysis of high school

students' *scientific reasoning* of grade 5–10 found weak and intermediate latent correlations of 0.33–0.73 between the subscales question, hypothesis, planning, and interpretation [33] (p. 56ff.). The author argued for a four-dimensional model. For this age group, the students were in the phase of acquiring knowledge on the phases that make up *scientific reasoning*. (iii) For the construct *experimentation competences*, manifest correlations of 0.38–0.74 were found for grade five and 0.64–0.78 for grade six between the three subscales of the SDDS model of Klahr [20]: search hypotheses, test hypotheses and evaluate evidence [34] (p. 42). While the study of Wellnitz with 10th graders suggested a 1D model, the other studies with younger students pointed out that experimentation related constructs require at least a two-dimensional model [33,34]. The phenomenon could be explained by a more integrative and interwoven processing of specialized knowledge coming along with study progress cf. [50]. More generally speaking, the fact of low latent correlations between the dimensions of *knowledge of what to assess* regarding experimentation competences was not surprising regarding the educational target groups within our study. A remarkable percentage of pre-service teachers, i.e., the undergraduates (53% of the sample), had not had very much biology teacher education courses (mean number of courses completed = 1.2) until their participation in our study.

For pre-service biology teachers, the subscales of *knowledge of what to assess* regarding experimentation competences showed the lowest latent correlations between hypothesis formation and analysis of data (0.57). In contrast, these two subscales correlated highest in studies conducted with high school students (correlation: 0.73 and 0.78 [33] (p. 58) [34] (p. 42)). For these students, the correlations between hypotheses and interpretation or between search hypotheses and evaluate evidence are explained by a greater relevance of domain-specific knowledge and less relevance of methodological knowledge in comparison to the phase planning/testing hypothesis [34] (p. 45). In our study with pre-service biology teachers, hardly no additional knowledge of biological phenomena was required to solve the tasks. For instance, a scenario included the information that the amount of released gas bubbles indicates the rate of photosynthesis of waterweed in the experiment. It was only required for pre-service teachers to know that oxygen is a product of photosynthesis to link a greater amount of gas bubbles to a greater rate of photosynthesis to interpret data given in the following scenario. Learning that oxygen is a product of photosynthesis is the content of school curricula for grade seven/eight [51]. Therefore, in our study, knowledge of biological phenomena should not have influenced the test results in contrast to the studies investigating high school students. The highest correlation of subscales of *knowledge of what to assess* regarding

experimentation competences existed between the design of an experiment and analysis of data (0.80), which was analogous to findings of Wellnitz's study of *scientific inquiry* [22] (p. 132). This could result from a stronger focus on these phases in research studies cf. [52,53] and perhaps as a consequence of teaching at university.

Regarding test quality, the 20 final items of *knowledge of what to assess* regarding experimentation competences in biology had a satisfactory item fit and discrimination for the 1D as well as the 3D model—except for the discrimination of Item 6 (0.15) and Item 18 (0.21). The low discrimination of items 6 and 18 could be explained by their great difficulty [54]. Only a few students solved both items. According to the contents of these items, we were interested in the knowledge of the term engineering mode of experimentation and its description (6). Second, to correct a hypothesis that is unfounded turned out to be a challenge. In biology teaching, there is a lack of clear rules concerning how to justify hypotheses. Several approaches exist that range from not addressing the fact that a hypothesis should be well-founded to expecting that a reason is given for the hypothesis [55,56]. Up to now, the issue of backing up hypotheses is not focused coherently in textbooks for school or teacher education [55–57].

The accuracy of the estimated item difficulties of the IRT analyses was given by the high item separation reliability of 0.99. The EAP/PV value, indicating the accuracy of the estimated person abilities, of 0.60 was comparable to tests measuring similar constructs. For example, a study measuring *scientific inquiry* (observing, comparing, experimentation) with 116 items [22] (p. 129) reached an EAP/PV reliability of 0.59 for high school students. Thereby, the subscale experimentation (22 items) reached a reliability of 0.41 (EAP/PV) in a 3D model with observing (0.37, 18 items) and comparing (0.39, 10 items) [22] (p. 136f.). Similar results were reached for an instrument measuring *scientific reasoning* with 24 items (EAP/PV = 0.69 (study I) and 0.68 (II) [33] (p. 51, 53)) and an instrument measuring *diagnostic competence* for students' experimentation competences with 17 items (Cronbach's alpha = 0.50, [14] (p. 189)). The measurement of a construct with full content and a limited number of items is in line with reduced reliability [58]. Since our construct *knowledge of what to assess* regarding experimentation competences covers the three phases of experimentation, the broad approach is reflected in the reliability of the instrument.

Moreover, low variances and open-answer formats can contribute to lower reliability [58,59]. On the upside, open tasks can measure skills closer to real-life performance than multiple-choice items and provide additional information [60].

In our study, some items were too difficult. While providing valuable information about *knowledge of what to assess*, they were not beneficial for precise measurement. Excluding difficult items or collapsing item steps could improve the quality of the instrument. More items for low and intermediate person abilities would improve the accuracy of the measurement [45] (p. 125f.). The low variances could result from a relatively homogenous sample of test persons (i.e., pre-service biology teachers).

Furthermore, in 2014/2015, *knowledge of what to assess* regarding experimentation competences might have hardly been addressed in teacher education courses, which is line with the lack of connecting CK, PCK, and PK (pedagogical knowledge) in German teacher education in that time [61]. Only in the last four years have there been nationwide efforts to systematically link these three knowledge areas further to develop the quality of teacher education [61]. However, each university, funded by the German Federal Ministry of Education and Research within the "Qualitätsoffensive Lehrerbildung", could decide its priorities for further developing their teaching. Thus, only a few universities addressed linking CK and PCK concerning competences in science (i.e., Technische Universität Braunschweig).

The chosen test length seemed suitable for measurement. One indicator was the difficulties of the items in the last scenario: Item 15 focusing on incorrect data analysis was comparably easy. This indicated that no respondent fatigue occurred. In contrast, item 16 (item category i) focuses on the student conception confirmation bias, which could explain why this item is more difficult than item 15.

The measurement instrument could be applied to undergraduate and graduate students of biology education. Significant DIF in the 1D model could only be detected for one item (i.e., Item 7 dealing with the systematic variation of variables). This item is considerably easier for graduate students than for undergraduate students. The fact could be due to an imprecise measurement related to the phenomenon of very low item difficulty. In addition, the DIF could be plausibly explained by specific training of the control of variables in (a) session(s) of teacher education courses, which was likely because the systematic variation of variables was one of the highlighted issues in reputable textbooks for German biology teacher education (e.g., [18]). In sum, the instrument was suitable to get an insight into pre-service teachers' *knowledge of what to assess* regarding experimentation competences in biology.

6.4.2 Validation

Latent correlations between the three constructs *knowledge of what to assess* regarding experimentation competences, *diagnostic competence* [14], and *examining competence* [35] were examined. *Knowledge of what to assess* regarding experimentation competences correlated highest with *examining competence* (0.78). The high correlation indicated a shared knowledge base. Both tests required knowledge about criteria for hypothesis formation, design of an experiment, and the analysis of data. Unexpectedly, the latent correlation between *knowledge of what to assess* regarding experimentation competences and *diagnostic competence* was comparably low (0.37). This could result from the test design. To solve the *diagnostic competence* tasks, the criteria for experimentation did not have to be known by the pre-service teachers. They were given, and pre-service teachers only had to identify whether they were fulfilled or not.

Moreover, the three instruments placed different emphasis on the successive phases of experimentation. Our instrument placed equal emphasis on the three phases hypothesis formation (seven items, 35% of items), design of an experiment (seven items, 35% of items), and analysis of data (six items, 30% of items). The instrument *diagnostic competence* had 16.6% of items focusing on hypothesis formation and 16.6% on the analysis of data. The majority, 67% of the items, dealt with the design of an experiment (42%) and performance in the sense of documentation (25%). The instrument for *examining competence* included the phase question formation, considering all four phases equally with 25% of the items. Considering the item distribution to the phases of experimentation in the three instruments investigated, the instrument for *examining competence*, and our instrument had a more similar emphasis on the different phases than the instrument for *diagnostic competence* and our instrument. The results have to be treated carefully due to the available instruments for related constructs for validation whose reliabilities are improvable.

The finding that only advanced students' *self-efficacy beliefs* correlated with *knowledge of what to assess* regarding experimentation competences could be explained by a better understanding of the contents addressed in the *self-efficacy beliefs* instrument by advanced students. During their studies, they engage with these topics and, consequently, they could achieve a more accurate ability to report on their *self-efficacy* regarding these subscales. Correlations of *knowledge of what to assess* with *self-efficacy beliefs* regarding *planning and conducting lessons in consideration of research results on biology education* and *planning lessons in consideration of core concepts and competences*

for biology [42] were an indicator for validity since both *self-efficacy* subscales comprised information relevant for the assessment of experimentation competences.

As assumed, the number of learning opportunities and the performance in high school biology as well as biology courses and biology teacher education courses at university (grades) correlated with person abilities. This finding indicated that the test measured knowledge and skills acquired at university. The average grade in courses in biology teacher education showed a higher correlation with *knowledge of what to assess* regarding experimentation competences than the average grade in courses in biology at university, which could be explained by the higher portion of biology school curricula procedural competences and contents used in the present study. The biology teacher education curriculum reflects the previously mentioned school curricula requirements to a certain degree [12].

The comparison of student abilities of students at the undergraduate and graduate levels showed higher person abilities for students at the graduate level, which was in accordance with our hypothesis regarding research question two and thus an indicator for validity. It underlined that the instrument measured knowledge that could probably be acquired during biology teacher education.

6.4.3 Strengths and Weaknesses of Pre-Service Biology Teachers Regarding Knowledge of What to Assess Regarding Experimentation Competences

Person abilities of pre-service biology teachers of *knowledge of what to assess* did not differ significantly for the three phases of experimentation.

Studies of high school students' *scientific reasoning* and *scientific inquiry* found that the interpretation of data analysis was more straightforward than the formation of hypotheses ([33] (p. 63) (grade 5–10) [22] (p. 141) (grade 10)). The findings for the phase design of an experiment were diverse and ranged from most difficult in some studies ([34] (p. 41) (grade 5 and 6); [33] (p. 63) to easiest in another [22] (p. 141)). The operationalization of the constructs could influence the results.

The finding of similar difficulties of the assessment of the three phases in our project was in line with skills pre-service teachers were expected to possess or acquire in their education. Having to teach and assess the whole process of experimentation, no significant differences in difficulties regarding the three phases should occur. However, specific criteria for experimentation proved to be challenging to assess, such as the founded hypothesis

(Item 3, 5, 18). This criterion might not have been trained explicitly and intensively at school and university, which made it difficult to solve the tasks of the test instrument.

The restricted knowledge of scientific terminology by pre-service biology teachers in our study was striking. It could have been caused by a certain lack of precise communicative skills in the teacher education curriculum [12] and thus probably limited course time spent on teaching and practicing scientific terms. Furthermore, the study provided hints that misconceptions concerning experimentation competences were hard to identify for pre-service teachers, which could be explained by the fact that experimentation competences can benefit from different sources, such as CK taught in natural science subjects as well as from PCK taught in teacher education courses. Instead, student misconceptions were mainly taught about in PCK related teacher education courses. Comparing the portions of CK and PCK in the biology teacher education curriculum for secondary school teachers (that made up the most significant part of our participants), the share of PCK was much smaller than the share of CK [12]. In addition, the assessment of correct (item category ii) and incorrect student solutions (item category iv) was differently demanding. Correct student solutions in our study were a lot more challenging to assess than incorrect student solutions. Analogous to more complex features of compensatory decision-making in comparison with non-compensatory decision-making [62], the consideration of positive as well as negative aspects in student performance for an assessment was more demanding and consequently more difficult than concentrating on a mistake or disadvantage only.

Moreover, the assessment of specific criteria concerning the planning of the standardization and accuracy of the performance and measurement (item category iii) was very demanding. Despite dealing with incorrect student solutions, Item 10 (lack of standardization) and Item 14 (inaccurate measurement procedure) were difficult. The results could be explained by the neglect of these criteria in the curricula [51]. This was also reflected in the findings of high school students' (grade 12) experimentation competences. Less than 22% of high school students considered when and how often to perform measurements during the planning of an experiment [63].

Thus, the present study gave insights into which aspects of PCK relevant knowledge and skills concerning *knowledge of what to assess* regarding experimentation competences are already well taught and learned. In addition, it revealed the remaining challenges for further developing biology teacher education.

6.4.4 Limitations

The comparison of the 1D and 3D models of *knowledge of what to assess* regarding experimentation competences did not provide a clear result regarding the dimensionality of the construct. Higher latent correlations were expected for the 1D model. An analysis with more items per subscale—so that all parts of the three subscales are better covered—could shed a brighter light on the question of dimensionality in order to evaluate how far the assessment of the three phases of experimentation requires similar knowledge and skills.

Considering the SDDS model [20], our construct *knowledge of what to assess* regarding experimentation competences included the assessment of experimentation competences regarding the three phases hypothesis formation, design of an experiment, and analysis of data. The formation of questions and performance of experiments that are part of some models or conceptualizations of experimentation competences cf. [64] were not considered.

In the assessment tasks, we worked with descriptions of biology classroom scenarios close to reality. Instead of this, videos of students experimenting in the classroom could measure pre-service teachers' *knowledge of what to assess* closer to reality. More comprehensive tasks would reduce the number of tasks required. At the same time, it could increase the quality of information gained regarding the knowledge measured. However, a more realistic (and more complex) assessment situation could divert the focus from the experimentation competences, which has carefully been weighed against a more focused assessment situation with reduced complexity, as applied in our study.

6.5 Conclusions

Knowledge of what to assess regarding experimentation competences could be modeled and measured reliably and validly. The analyzed data comprised assessments of 495 pre-service teachers of seven German federal states and 18 universities. The database included pre-service biology teachers of different semesters, different study programs, and different school types. Thus, the following conclusions are drawn more generally for biology teacher education. Further studies could shed light on certain pre-service biology teacher subsamples more specifically.

We worked out criteria for the assessment of experimentation competences regarding hypothesis formation, design of an experiment, and analysis of data according to the SDDS model [20]. With this approach, we gained knowledge for evidence-based biology

teacher education in the field of teaching experimentation competences. The assessment tasks regarding these experimental phases of the developed instrument and the scenario format can—with adaptations based on the evidence given in our study—be used for teacher education designing teaching and learning environments to foster teaching experimentation competences in pre-service teachers. Thus, using the seven scenarios and not exceeding 90 minutes in testing time was an adequate approach.

Our study gave insights into pre-service teachers' strengths and weaknesses in the assessment of experimentation competences. The difficulties could be explained about the tasks. Assessing student conceptions as well as correct student solutions, turned out to be more difficult than assessing incorrect student solutions most of the time. The only exceptions we found concern the planning of a standardized and accurate performance and measurement. Comparably few pre-service teachers mastered these requirements. The results suggest that even more attention could be paid in teacher education on student conceptions to enable relevant assessments to be able to foster student learning systematically. Moreover, the relevance of knowing and understanding PCK relevant scientific terms for precise assessments should be highlighted in biology teacher education.

Our study on *knowledge of what to assess* focused on one of the four areas of assessment knowledge and skills defined by Abell and Siegel [16]. Further research could examine the other areas of knowledge about the assessment of experimentation competences as well as the relationships among the knowledge areas. Moreover, an examination of the relation of assessment literacy and the ability to plan lessons under consideration of students' experimentation competences, which is closely linked to "knowledge of assessment interpretation and action-taking", can give insights for improving teaching experimental lessons [16] (p. 215). For this purpose, we have an overlapping sample with a parallel study on teaching competences for experimental lessons, with one focus on the ability to plan lessons [65] that needs to be analyzed in the future.

The 1D model makes sense regarding the interdependent complex assessment of experimentation competences. Nevertheless, having a closer look at (i) the three phases of experimentation that have to be assessed and (ii) by grouping the items by the specific challenges that have to be overcome provides more in-depth insights into pre-service teachers' strengths and weaknesses. Thereby, the study clearly shows what teacher education already tackles to a good extent and what could be more addressed in the future to bring forward pre-service teachers' *knowledge of what to assess*: This helps to reflect and further develop current practices in biology teacher education in the field of improving

student experimentation competences. At the same time, it motivates further research to improve biology teacher education to overcome assessment challenges and to foster assessment literacy.

Author Contributions: Conceptualization, C.J. and S.B.; Data curation, C.J.; Formal analysis, C.J.; Funding acquisition, S.B. and M.H.; Investigation, C.J., S.B., M.H. and C.H.C.; Methodology, C.J., S.B. and C.H.C.; Project administration, S.B.; Resources, S.B.; Supervision, S.B. and M.H.; Validation, C.J. and S.B.; Visualization, C.J. and S.B.; Writing—original draft, C.J.; Writing—review & editing, S.B. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01PK11014B.

Acknowledgments: We thank the KoKoHs team, all pre-service biology teachers who participated in our study and the university staff who contributed to conduct the investigation.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

Table 6-10 Item content and scoring, items in order of decreasing difficulty per scoring category (H: Hypothesis formation, D: Design, and performance of the experiment, A: Analysis of data)

Item	Item 6	Item 4	Item 14	Item 15	Item 16	Item 10	Item 11	Item 13
Phase	D	H	D	A	A	D	A	
Criterion	Student conception: *Engineering mode*	*Untestable hypothesis*	*Inaccurate measurement procedure*	*Incorrect data analysis*	Student conception: *Confirmation bias*	*Lack of standardization*	*Imprecise data analysis*	*A conclusion without a link to the hypothesis*
The specific requirement for a full credit for the item	Naming and explaining				Explaining	Naming	Naming	
	the student conception "engineering mode"	that the hypothesis is untestable	that the measurement procedure is inaccurate	that the data analysis is incorrect	the student conception "confirmation bias"	that several aspects lack standardization	that not all findings are taken account of	that the conclusion is not linked to the hypothesis
Score 2	The criterion is named and explained				The criterion is explained or named completely		The criterion is named precisely	
Score 1	The criterion is named or for Item 4 and 6 explained				The criterion is explained or named in parts		The criterion is named imprecisely	
Score 0	The criterion is neither named nor explained							

Table 6-10 *Cont.*

Item	Item 17	Item 18	Item 19	Item 20
Phase	H		D	A
Criterion	*Untestable hypothesis*	*Unfounded hypothesis*	*Unsystematic variation of variables*	*A conclusion without a link to the hypothesis*
The specific requirement for a full credit for the item	Correcting the given insufficient student answer by stating			
	a testable hypothesis	a founded hypothesis	a systematic variation of variables	a conclusion that supports the hypothesis
Score 2	Student answer is corrected completely			
Score 1	Student answer is corrected in parts			Student answer is corrected in parts
Score 0	Student answer is not corrected			

Item	Item 3	Item 5	Item 2	Item 1	Item 8	Item 9	Item 7	Item 12
Phase	H		H		D			A
Criterion	*Founded hypothesis*	*Unfounded hypothesis*	*Testable hypothesis*	*Incomprehensive hypothesis formation*	*Systematic variation of variables*	*Imprecise design of experiment*	*Unsystematic variation of variables*	*Missing error analysis*
The specific requirement for a full credit for the item	Naming		Naming or describing					
	that the hypothesis is founded	that the hypothesis is unfounded	that the hypothesis is testable	that the hypothesis formation is incomprehensive	that the variation of variables is systematic	that the design of the experiment is imprecise	that the variation of variables is unsystematic	that error analysis is missing
Score 2	The criterion is named		The criterion is named or described					
Score 1								
Score 0	The criterion is neither named nor described							

References

1. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ.* **1998**, *5*, 7–74. [CrossRef]
2. Abell, S.K.; Volkmann, M.J. *Seamless Assessment in Science: A Guide for Elementary and Middle School Teachers*; Heinemann: Portsmouth, NH, USA, 2006; ISBN 978-0-325-00769-4.
3. Hattie, J.; Jaeger, R. Assessment and Classroom Learning: A deductive approach. *Assess. Educ.* **1998**, *5*, 111–122. [CrossRef]
4. Kultusministerkonferenz, K.M.K. (Ed.) *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*; Wolters Kluwer Deutschland GmbH: München, Germany, 2005.
5. National Research Council. *National Science Education Standards*; The National Academies Press: Washington, DC, USA, 1996; ISBN 978-0-309-05326-6.
6. Hammann, M. Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung—Dargestellt anhand von Kompetenzen beim Experimentieren. *MNU* **2004**, *57*, 196–203.
7. Hammann, M.; Phan, T.T.H.; Ehmer, M.; Bayrhuber, H. Fehlerfrei Experimentieren. *MNU* **2006**, *59*, 292–299.
8. Xu, Y.; Brown, G.T.L. Teacher assessment literacy in practice: A reconceptualization. *Teach. Teach. Educ.* **2016**,*58*, 149–162. [CrossRef]
9. Klieme, E.; Avenarius, H.; Blum, W.; Döbrich, P.; Gruber, H.; Prenzel, M.; Reiss, K.; Riquarts, K.; Rost, J.; Tenorth, H.-E.; et al. *The Development of National Education Standards: An Expertise*; Bundesministerium für Bildung und Forschung: Berlin, Germany, 2004.
10. Winterton, J.; Delamare-Le Deist, F.; Stringfellow, E. *Typology of Knowledge, Skills, and Competences: Clarification of the Concept and Prototype*; Office for Official Publications of the European Communities: Luxembourg, 2006; ISBN 92-896-0427-1.
11. Méhaut, P.; Winch, C. The European Qualification Framework: Skills, Competences or Knowledge? *Eur. Educ. Res. J.* **2012**, *11*, 369–381. [CrossRef]
12. der Kultusministerkonferenz, B. (Ed.) Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Available online: https: //www.kmk. org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (accessed on 16 March 2020).
13. Stiggins, R. Assessment Literacy. *Phi Delta Kappan* **1991**, *72*, 534–539.
14. Dübbelde, G. Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. Ph.D. Thesis, Universität Kassel, Kassel, Germany, 2013. Available online: https: //kobra.uni-kassel.de/handle/123456789/2013122044701 (accessed on 16 March 2020).
15. Magnusson, S.; Krajcik, J.; Borko, H. Nature, Sources and Development of Pedagogical Content Knowledge for Science Teaching. In *Examining Pedagogical Content Knowledge*; Gess-Newsome, J., Lederman, N.G., Eds.; Springer: Dordrecht, The Netherlands, 1999; pp. 95–132. ISBN 978-0-7923-5903-6.

16. Abell, S.K.; Siegel, M.A. Assessment Literacy: What Science Teachers Need to Know and Be Able to Do. In *The Professional Knowledge Base of Science Teaching*; Corrigan, D., Dillon, J., Gunstone, R., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp. 205–221. ISBN 978-90-481-3926-2.
17. OECD. PISA for Development Science Framework. In *PISA for Development Assessment and Analytical Framework: Reading, Mathematics and Science*; OECD Publishing: Paris, France, 2018; pp. 71–97. [CrossRef]
18. Schulz, A.; Wirtz, M.; Starauschek, E. Das Experiment in den Naturwissenschaften. In *Experimentieren im Mathematisch-Naturwissenschaftlichen Unterricht*; Rieß, W., Wirtz, M., Barzel, B., Schulz, A., Eds.; Waxmann: Münster, Germany, 2012; pp. 15–18.
19. Klautke, S. Ist das Experimentieren im Biologieunterricht noch zeitgemäß? *MNU* **1997**, *50*, 323–329.
20. Klahr, D. *Exploring Science: The Cognition and Development of Discovery Processes*; The MIT Press: Cambridge, MA, USA, 2000.
21. Li, J.; Klahr, D. The Psychology of Scientific Thinking: Implications for Science Teaching and Learning. In *Teaching Science in the 21st Century*; Rhoton, J., Shane, P., Eds.; NSTA Press: Arlington, VA, USA, 2006; pp. 307–328.
22. Wellnitz, N. *Kompetenzstruktur und -Niveaus von Methoden Naturwissenschaftlicher Erkenntnisgewinnung*; Logos: Berlin, Germany, 2012.
23. Ehmer, M. Förderung von kognitiven Fähigkeiten beim Experimentieren im Biologieunterricht der 6. Klasse: Eine Untersuchung zur Wirksamkeit von methodischem, epistemologischem und negativem Wissen. Ph.D. Thesis, Christian-Albrechts-Universität Kiel, Kiel, Germany, 2008. Available online: https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00002469/diss_ehmer.pdf (accessed on 16 March 2020).
24. Koslowski, B. *Theory and Evidence: The Development of Scientific Reasoning*; The MIT Press: Cambridge, MA, USA, 1996.
25. Mayer, J.; Ziemek, H.-P. Offenes Experimentieren: Forschendes Lernen im Biologieunterricht. *Unterr. Biol.* **2006**, *317*, 4–12.
26. Krüger, D. Bezaubernde Biologie—Mit Hypothesen der Lösung auf der Spur. *MNU* **2009**, *62*, 41–46.
27. Schauble, L.; Klopfer, E.; Raghaven, K. Students' Transition from an Engineering Model to a Science Model of Experimentation. *J. Res. Sci. Teach.* **1991**, *9*, 859–882. [CrossRef]
28. Köhler, K. Welche fachgemäßen Arbeitsweisen werden im Biologieunterricht eingesetzt? In *Biologie Didaktik. Praxishandbuch für die Sekundarstufe I und II*; Spörhase-Eichmann, U., Ruppert, W., Eds.; Cornelsen: Berlin, Germany, 2004; pp. 146–159.
29. Straube, P. *Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik*; Logos: Berlin, Germany, 2016.
30. Bouffard-Bouchard, T.; Parent, S.; Larivee, S. Influence of Self-Efficacy on Self-Regulation and Performance among Junior and Senior High-School Age Students. *Int. J. Behav. Dev.* **1991**, *14*, 153–164. [CrossRef]
31. Tschannen-Moran, M.; Woolfolk Hoy, A.; Hoy, W.K. Teacher Efficacy: Its Meaning and Measure. *Rev. Educ. Res. 1998, 68, 202-248* [CrossRef]
32. Saklofske, D.; Michaluk, B.; Randhawa, B. Teachers' Efficacy and Teaching Behaviors. *Psychol. Rep.* **1988**, *63*, 407–414. [CrossRef]

33. Grube, C.R. Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung: Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I. Ph.D. Thesis, Universität Kassel, Kassel, Germany, 2010. Available online: https: //kobra.uni-kassel.de/handle/ 123456789/2011041537247 (accessed on 16 March 2020).
34. Hammann, M.; Phan, T.T.H.; Bayrhuber, H. Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? *Z. Erzieh.* **2007**, *8*, 33–49.
35. Krüger, D.; Upmeier zu Belzen, A.; Nordmeier, V.; Tiemann, R.; Hartmann, S.; Mathesius, S.; Stiller, J.; Straube, P. Kooperation der Projekte Ko-WADiS und ExMo. Unpublished.
36. Bögeholz, S.; Joachim, C.; Hasse, S.; Hammann, M. Kompetenzen von (angehenden) Biologielehrkräften zur Beurteilung von Experimentierkompetenzen. *Unterrichtswissenschaft* **2016**, *44*, 40–54.
37. List, M.K. Testbearbeitungsverhalten in Leistungstests: Modellierung von Testabbruch und Leistungsabfall. Ph.D. Thesis, Christian-Albrechts-Universität Kiel, Kiel, Germany, 2018. Available online: https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00007735/diss_mk_list_testbearbeitungsverhalten_in_leistungstests.pdf (accessed on 16 March 2020).
38. Kultusministerium, N. (Ed.) Kerncurriculum für das Gymnasium Schuljahrgänge 5 -10: Naturwissenschaften. 2007. Available online: http: //db2.nibis.de/1db/cuvo/datei/kc_gym_nws_07_nib.pdf (accessed on 16 March 2020).
39. Mayring, P. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*; Beltz: Weinheim, Germany, 2010.
40. Donner, A.; Rotondi, M.A. Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters. *Int. J. Biostat.* **2010**, *6*. [CrossRef]
41. De Swert, K. Calculating Inter-Coder Reliability in Media Content Analysis Using Krippendorff's Alpha. Available online: https: //www.polcomm.org/wp-content/uploads/ICR01022012.pdf (accessed on 16 March 2020).
42. Mahler, H. Selbstwirksamkeitserwartungen angehender Biologielehrkräfte—Entwicklung eines Messinstrumentes. Master's Thesis, Georg-August-Universität Göttingen, Göttingen, Germany, 2014. Unpublished.
43. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174. [CrossRef]
44. Wu, M.L.; Adams, R.J.; Wilson, M.R.; Haldane, S.A. *ACER ConQuest Version 2.0: Generalised Item Response Modelling Software*; Australian Council for Educational Research: Camberwell, Victoria, Australia, 2007.
45. Boone, W.J.; Staver, J.R.; Yale, M.S. *Rasch Analysis in the Human Sciences*; Springer: Dordrecht, The Netherlands, 2014.
46. Bond, T.G.; Fox, C.M. *Applying the Rasch Model*; Routledge: New York, NY, USA, 2015. 47.
47. OECD. *PISA 2006 Technical Report*; OECD: Paris, France, 2009.
48. Pohl, S.; Carstensen, C.H. *NEPS Technical Report—Scaling the Data of the Competence Tests (NEPS Working Paper No. 14)*; Otto-Friedrich-Universität, Nationales Bildungspanel: Bamberg, Germany, 2012.
49. Schecker, H.; Parchmann, I. Modellierung naturwissenschaftlicher Kompetenz. *ZfDN* **2006**, *12*, 45–66.

50. Velten, S.; Nitzschke, A.; Nickolaus, R.; Walker, F. Die Fachkompetenzstruktur von Technikern für Elektrotechnik und Einflussfaktoren auf ihre Kompetenzentwicklung. *J. Technol. Educ.* **2018**, *6*, 201–222.
51. Kultusministerium, N. (Ed.) Kerncurriculum für das Gymnasium Schuljahrgänge 5–10: Naturwissenschaften. 2015. Available online: https://db2.nibis.de/1db/cuvo/datei/nw_gym_si_kc_druck.pdf (accessed on 16 March 2020).
52. Völzke, K.; Arnold, J.; Kremer, K. Denken und Verstehen beim naturwissenschaftlichen Problemlösen. Eine explorative Studie. *Z. Interpret. Schul Unterr.* **2013**, *2*, 58–86. [CrossRef]
53. Chen, Z.; Klahr, D. All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Dev.* **1999**, *70*, 1098–1120. [CrossRef] [PubMed]
54. University of Washington. Understanding Item Analyses. Available online: https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/ (accessed on 16 March 2020).
55. Baack, K.; Steinert, K. *Natura 7/8 Biologie für Gymnasien, Niedersachsen*; Klett: Stuttgart, Germany, 2015.
56. Hammann, M. Experimentieren. In *Biologie-Methodik. Handbuch für die Sekundarstufe I und II*; Spörhase, U., Ruppert, W., Eds.; Cornelsen: Berlin, Germany, 2014; pp. 102–106.
57. Bayrhuber, H.; Hammann, M. (Eds.) *Linder Biologie: Abi-Aufgabentrainer, Wissen Anwenden und Kompetenzen Einüben*; Schroedel: Braunschweig, Germany, 2013.
58. Bühner, M. *Einführung in die Test- und Fragebogenkonstruktion*; Pearson: Hallbergmoos, Germany, 2011.
59. Stecher, B.M.; Klein, S.P. The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educ. Eval. Policy Anal.* **1997**, *10*, 1–14. [CrossRef]
60. Shavelson, R.F. *Measuring College Learning Responsibly: Accountability in a New Era*; Stanford University Press: Stanford, CA, USA, 2009.
61. Bundesministerium für Bildung und Forschung. Qualitätsoffensive Lehrerbildung. Available online: https://www.qualitaetsoffensive-lehrerbildung.de/de/fachwissenschaften-fachdidaktik-und- bildungswissenschaften-1803.html (accessed on 10 February 2020).
62. Eggert, S.; Bögeholz, S. Students' Use of Decision-Making Strategies With Regard to Socioscientific Issues: An Application of the Rasch Partial Credit Model. *Sci. Educ.* **2010**, *94*, 230–258. [CrossRef]
63. Arnold, J.; Kremer, K.; Mayer, J. Wissenschaftliches Denken beim Experimentieren – Kompetenzdiagnose in der Sekundarstufe II. *Erkenn. Biol.* **2012**, *11*, 7–20.
64. Mayer, J.; Grube, C.; Möller, A. Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In *Lehr- und Lernforschung in der Biologiedidaktik (Band 3)*; Harms, U., Sandmann, A., Eds.; StudienVerlag: Innsbruck, Austria, 2008; pp. 63–78.
65. Hasse, S.; Joachim, C.; Bögeholz, S.; Hammann, M. Assessing teaching and assessment competences of biology teacher trainees: Lessons from item development. *Int. J. Educ. Math. Sci. Technol.* **2014**, *2*, 191–205. [CrossRef]

7. Zusammenfassung und Diskussion

Diese Arbeit untersucht Wissen und Fähigkeiten über die angehende Biologielehrkräfte verfügen sollten, um Experimentierkompetenzen von Schüler*innen im Unterricht angemessen beurteilen zu können – die Basis für adaptive Förderung. Die Arbeit fokussiert dabei die besonders fachspezifische Wissenskategorie *knowledge of what to assess* von *assessment literacy* (Abell & Siegel, 2011) im Kontext Experimentierkompetenzen von Schüler*innen unterrichtlich fördern. Die Bedeutung von Wissen und Fähigkeiten für (angehende) Lehrkräfte ergibt sich aus dem Einfluss von Lehrkräften auf das Lernen der Schüler*innen (Brunner et al., 2011; Ruiz-Primo & Furtak, 2007). Die Relevanz von Experimentierkompetenzen für Schüler*innen ist begründet durch die zentrale Rolle der experimentellen Methode für die Erkenntnisgewinnung in der Biologie (Klautke, 1997, S. 323; Kultusministerkonferenz, 2005; Schulz et al., 2012, S. 15). Grundlegendes Wissen für Beurteilungen und die Anwendung des Wissens zur Beurteilung sind Inhalte der Ausbildung von Lehrkräften (Kultusministerkonferenz, 2019b). Die Vorgaben für die Lehrerbildung geben Wissen für die fachbezogene Leistungsbeurteilung allerdings nur allgemein als Ziel vor und schlüsseln Wissen für Beurteilungen z.B. nicht für einzelne Kompetenzbereiche auf (Kultusministerkonferenz, 2019b, S. 22).

Zur Zeit des ExMo Projekts lagen in der Lehrerbildungsforschung für den bundesdeutschen Kontext nur wenige Erkenntnisse darüber vor, inwiefern angehende Biologielehrkräfte über das notwendige Wissen verfügen (Abell & Siegel, 2011, S. 207). Ziel dieser Arbeit war es, *knowledge of what to assess* in Bezug auf Experimentierkompetenzen zu modellieren und zu messen. Eine wichtige Voraussetzung dafür war ein geeignetes Messinstrument. Auf Basis des theoretisch hergeleiteten Konstrukts *knowledge of what to assess* in Bezug auf Experimentierkompetenzen (in Kapitel 4 als *assessment competence* und in Kapitel 5 als *Beurteilungskompetenz* benannt) wurde ein Rahmenkonzept für die Entwicklung und Auswertung von Aufgaben für das Messinstrument entwickelt. Die Entwicklung der Aufgaben umfasste mehrere Schritte. Die Reliabilität und Validität wurden in der Pilotstudie (N=145) (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016) und Hauptstudie (N=500) (vgl. Kapitel 6, Joachim et al., 2020) sowie mittels einer Expertenbefragung (vgl. Kapitel 4, Hasse et al., 2014) untersucht. Anhand der Hauptstudiendaten wurde die Dimensionalität des Konstrukts systematisch analysiert und Stärken und Schwächen von Biologielehramtsstudierenden mit Blick auf *knowledge of what to assess* in Bezug auf Experimentierkompetenzen in den Blick genommen.

Vor dem theoretischen Hintergrund zu Wissen und Fähigkeiten für die Beurteilung von Experimentierkompetenzen (Kapitel 2) werden die Vorgehensweise und Ergebnisse der Arbeit zusammengefasst und diskutiert. Entsprechend der Ziele und Forschungsfragen (Kapitel 3) der Arbeit wird zunächst die Entwicklung und Eignung des Instruments zur Erfassung von *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Experimentierkompetenzen zusammenfassend dargestellt und diskutiert (Kapitel 7.1). Anschließend werden die empirischen Ergebnisse zur Dimensionalität (Kapitel 7.2) sowie Hinweise auf die Validität (Kapitel 7.3) zusammengefasst und diskutiert. Schließlich werden Erkenntnisse zu Stärken und Schwächen von Biologielehramtsstudierenden beim Beurteilen von Experimentierkompetenzen zusammenfassend beschrieben und diskutiert (Kapitel 7.4). Im abschließenden Kapitel der Arbeit wird ein Fazit gezogen und weitere Forschungsperspektiven aufgezeigt (Kapitel 8).

7.1 Zusammenfassung und Diskussion zu Forschungsfrage (1)

Die Forschungsfrage (1) steht in Bezug zum Ziel der Entwicklung eines Instruments zur Erfassung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen. Sie lautet:

(1) Inwiefern kann ein Instrument zur reliablen Messung des Konstrukts *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Schülerexperimentierkompetenzen entwickelt werden?

Der Forschungsfrage (1) wurde in den Publikationen 1, 2 und 3 nachgegangen. Insgesamt kann festgehalten werden, dass ein Instrument zur Erfassung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen entwickelt wurde (vgl. Kapitel 6, Joachim et al., 2020). Die Entwicklung von Testaufgaben zur Erfassung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen erfolgte sehr systematisch auf Basis eines Rahmenkonzepts und schrittweise unter Durchführung mehrerer Vortestungen (Studie Lauten Denkens, N=16; Pre-Pilotierung, N=24) (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016), um z.B. die Verständlichkeit der Testaufgaben sicherzustellen.

Die Testaufgaben bestehen aus einem Unterrichtsszenario, in dem Experimentierkompetenzen von Schüler*innen dargestellt werden, und jeweils szenariobezogenen offenen Beurteilungsaufgaben. Das Design der Testaufgaben wird den Anforderungen kriteriengeleiteter Statusdiagnosen (u.a. indem Kriterien zum Experimentieren zur Beurteilung

vorliegender Experimentierkompetenzen hypothetischer Schüler*innen angewandt werden) gerecht, ohne durch wenig relevante Zusatzinformationen die Studierenden von einer kriteriengeleiteten Beurteilung abzulenken. Es ermöglicht eine Erfassung der Fähigkeit, Wissen für die Beurteilung von Experimentierkompetenzen anzuwenden. Mit den Aufgaben können qualitative Einblicke in *knowledge of what to assess* in Bezug auf Experimentierkompetenzen bei Studierenden gewonnen werden.

Als repräsentative, relevante Kontexte wurden Samenkeimung (Klassenstufe 5/6), Fotosynthese (Klassenstufe 7/8) und Enzymatik (Klassenstufe 9/10) (Niedersächsisches Kultusministerium, 2007) gewählt. Die Testaufgaben sind so konstruiert, dass kein fachinhaltliches Wissen zu den drei Kontexten für die Bearbeitung notwendig ist. In den Beurteilungsaufgaben wird methodisches Wissen bzw. Kompetenzen von Schüler*innen bezüglich der drei Phasen Hypothesenbildung, Planung von Experimenten oder Auswertung von Daten fokussiert (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016 und Kapitel 6, Joachim et al., 2020). Die Inhaltsvalidität der Testaufgaben wurde mittels einer Expertenbefragung (N=6) abgesichert.

Die empirische Weiterentwicklung des Instruments erfolgte schwerpunktmäßig in der Pilotstudie (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016). Die Publikation zur Hauptstudie enthält die finale Modellierung von *knowledge of what to assess* und das entsprechende Messinstrument (vgl. Kapitel 6, Joachim et al., 2020). In der Pilotstudie wurden 145 Biologielehramtsstudierende von acht deutschen Universitäten befragt. Insgesamt wurden 27 Aufgaben in einem *incomplete block design* in neun verschiedenen Testheften mit je neun Aufgaben eingesetzt. Jedes Testheft umfasste zu jedem der drei Kontexte Samenkeimung, Fotosynthese und Enzymatik jeweils eine Aufgabe zur Hypothesenbildung, eine zur Planung von Experimenten und eine zur Auswertung von Daten (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016).

Mit dem Ziel direkt vergleichbare und möglichst vollständige Daten zu den einzelnen Beurteilungsaufgaben zu gewinnen und für eine möglichst gute Stichprobengröße für eine verlässliche Anwendung von Kompetenzmodellierungen, wurde für die Hauptstudie – unter Mehrzieloptimierungsgesichtspunkten – nur ein Testheft erstellt. Die Bearbeitungszeit sollte inklusive der in der Hauptstudie eingesetzten Validierungsinstrumente zumutbare 90 Minuten nicht überschreiten, um nachlassende Konzentration bei der Bearbeitung der Aufgaben zu reduzieren bzw. zu vermeiden und eine möglichst hochwertige Bearbeitungsqualität zu erzielen (List, 2018). Auch ermöglicht die Testzeit von 90 Minuten die Befragung möglichst vieler Testpersonen im Rahmen von Seminaren bzw.

freien Zeitslots zwischen Lehrveranstaltungen. Folglich wurden für das Instrument der Hauptstudie sieben Testaufgaben aus der Pilotstudie ausgewählt, die sich besonders gut für die Erfassung des Wissens geeignet hatten und zusammen die Breite des Wissens zu Kriterien des Experimentierens und Vorstellungen von Schüler*innen zum Experimentieren abdecken sowie jede der drei Phasen des Experimentierens jeweils in mindestens zwei verschiedenen Kontexten (Samenkeimung, Fotosynthese oder Enzymatik) umfassen. Teilweise wurden die Beurteilungsaufgaben zu den Szenarien noch ergänzt, um auch mit sieben Aufgaben (im Vergleich zu neun Aufgaben in der Pilotstudie) *knowledge of what to assess* in Bezug auf Experimentierkompetenzen umfassend zu erfassen. Für die in Kapitel 6 (Joachim et al., 2020) dargestellte Hauptstudie wurden Daten von 500 Lehramtsstudierenden von 18 deutschen Universitäten erhoben.

Mit der Entwicklung von geeigneten Testaufgaben verbunden sind Überlegungen zur Auswertung von Antworten von angehenden Lehrkräften auf offene Beurteilungsaufgaben. Die Auswertung der Antworten der Pilotstudie und Hauptstudie berücksichtigt den Systematisierungs- und Elaborationsgrad der Antworten (vgl. z.B. Hammann, 2004; Upmeier zu Belzen & Krüger, 2010). Ein hoher Score wurde z.B. für die Nennung eines Schülerfehlers oder Kriteriums zum Experimentieren in Verbindung mit dessen Erklärung am Kontext vergeben. Das Scoringprinzip beruht auf der Annahme, dass die Nennung eines Schülerfehlers oder Kriteriums allein nicht dessen Verständnis nachweist. Eine korrekte kontextualisierte Erklärung indiziert, dass die genannten Vorstellungen und Kriterien hoch wahrscheinlich (weitgehend) verstanden wurden. Erklärungen wurden in den Beurteilungsaufgaben eingefordert. Für die Hauptstudie wurde das Scoring weiterentwickelt. Der Ansatz einer Vergabe höherer Scores für einen höheren Systematisierungs- und Elaborationsgrad wurde beibehalten (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016 und Kapitel 6; Joachim et al., 2020).

Die Daten der Pilotstudie wurden klassisch statistisch ausgewertet. Die Reliabilität war zufriedenstellend: Für sieben der neun Testhefte wurde ein Cronbachs alpha von mindestens 0.70 berechnet (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016).

Für die Hauptstudie (vgl. Kapitel 6, Joachim et al., 2020) wurde das Rasch Partial Credit Modell angewandt. Die Test- und Itemkennwerte des ein- und dreidimensionalen Modells (vgl. Kapitel 7.2) waren zufriedenstellend. Im Folgenden soll der Fokus auf dem eindimensionalen Modell liegen, auf Basis dessen die weiteren Analysen erfolgt sind. Die *item separation reliability* erreichte einen hohen Wert von 0.99. Die EAP/PV Reliabi-

lität von 0.60 ist vergleichbar mit Reliabilitäten von verwandten Konstrukten, z.B. *diagnostische Kompetenzen* und *naturwissenschaftliche Untersuchungen* (Dübbelde, 2013; Wellnitz, 2012). Sowohl die Breite des Instruments, die geringe Anzahl an Items des Instruments als auch die offenen Aufgaben sind denkbare Faktoren, die zu vergleichsweise geringen oder moderaten Reliabilitäten führen können (Bühner, 2011; Stecher & Klein, 1997). Die höheren Reliabilitäten der Testhefte der Pilotstudie (die EAP/PV Reliabilität ist in der Größenordnung vergleichbar mit Cronbachs alpha) sind ggf. zudem auf eine stärkere Reduktion der Items bei der Skalenkonstruktion (6-10 pro Testheft) auf die jeweils am besten geeigneten Items für die Skala zurückzuführen.

Die Item-Fit-Werte der Hauptstudie sind gut. Die geringe Varianz des Instruments kann auf eine immer noch vergleichsweise homogene Stichprobe zurückgeführt werden, handelt es sich doch ausschließlich um Biologielehramtsstudierende. *Differential Item Functioning* (DIF) war hinsichtlich des Studienabschnitts (Bachelor bzw. Staatsexamen ≤ Semester 6 und Master bzw. Stattsexamen ≥ Semester 7) mit Ausnahme von nur einem – plausibel erklärbarem – Item nicht gegeben. Durchgängig konnte zudem beim Vergleich der Geschlechter kein DIF festgestellt werden (vgl. Kapitel 6, Joachim et al., 2020). Knapp ein Drittel der dichotomen Items bzw. Itemstufen trichotomer Items werden jeweils von weniger als 16% der Biologielehramtsstudierenden gelöst. Aufgrund inhaltlicher Informationen zu *knowledge of what to assess* wurden diese Itemstufen und Items beibehalten. Eine Ergänzung von Items, deren Schwierigkeit den niedrigen und mittleren Personenfähigkeiten entspricht, könnte die Messgenauigkeit des Instruments verbessern (Boone et al., 2014, S. 125f.; vgl. Kapitel 6, Joachim et al., 2020).

Die gewählte Testlänge scheint gut geeignet für die Erfassung von *knowledge of what to assess* von Biologielehramtsstudierenden für Experimentierkompetenzen. Beobachtet werden konnte, dass die Schwierigkeiten der Items am Ende des Testhefts nicht höher liegen als die Schwierigkeiten von Items am Anfang oder der Mitte des Testhefts (vgl. Kapitel 6, Joachim et al., 2020).

Eine Grenze des Instruments ist die Konzentration auf die drei Phasen Hypothesenbildung, Planung von Experimenten und Auswertung von Daten durch Verwendung des SDDS-Modells (Klahr, 2000). *Knowledge of what to assess* in Bezug auf die Beurteilung von Kompetenzen Lernender naturwissenschaftliche Fragestellungen zu formulieren, wird damit nicht erfasst (vgl. Kapitel 6, Joachim et al., 2020).

Das offene Aufgabenformat in Kombination mit einer angemessenen Testzeit, um Konzentrations- oder Motivationsverluste zu reduzieren, erforderte eine begrenzte Anzahl

an Items. Die Modellierung basiert insgesamt auf 30 Items bzw. Itemstufen. In einem Multiple-Choice-Test könnten mehr Items eingesetzt werden. Die Antworten würden jedoch vermutlich weniger Einblicke in Wissen und Fähigkeiten für Beurteilungen ermöglichen als offene Antworten (Shavelson, 2009).

Die zu beurteilenden Schülerleistungen werden in unserem Instrument in Unterrichtsszenarien beschrieben. Diese gewählte und umgesetzte Form der Darstellung der Experimentierkompetenzen ermöglicht den angehenden Lehrkräften eine Fokussierung auf bestimmte Schülerkompetenzen sowie die Option, die dargestellten Schülerleistungen mehrfach und in Ruhe zu prüfen. Dadurch werden die Aufgaben dem Wissen und den Fähigkeiten angehender Lehrkräfte gerecht. Der Einsatz von Videosequenzen könnte eine realitätsnähere Erfassung ermöglichen sowie durch umfassendere Beurteilungsaufgaben weitere qualitative Erkenntnisse zu Wissen und Fähigkeiten Biologielehramtsstudierender liefern. Die größere Komplexität der dargestellten Situationen kann jedoch den Fokus auf die Beurteilung der Experimentierkompetenzen für Biologielehramtsstudierende erschweren, sodass Realitätsnähe und Fokussierung der Beurteilung jeweils gegeneinander abgewogen werden müssen (vgl. Kapitel 6, Joachim et al., 2020). Zu berücksichtigen bleibt in jedem Fall, dass Beurteilungen in der Praxis komplexer sind, z.B. weil äußere Faktoren, wie Bedürfnisse und Interessen von Schuladministrator*innen, Kolleg*innen sowie Eltern und Schüler*innen, dazu kommen und Kompromisse zwischen dem Wissen und den Vorstellungen von Beurteilungen und äußeren Faktoren gefunden werden müssen (Xu & Brown, 2016, S. 157). Zudem muss eine Lehrkraft im Unterricht vielen weiteren Aufgaben nachkommen (Kultusministerkonferenz, 2019a).

Zusammenfassend kann das in der Hauptstudie eingesetzte Instrument als geeignet für die Erfassung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen angesehen werden. Die umfangreiche Entwicklungsarbeit hat ein Instrument hervorgebracht, das in motivierender Weise *knowledge of what to assess* realitätsnah und in relevanten Kontexten der Sekundarstufe I erfasst (vgl. Kapitel 6, Joachim et al., 2020).

7.2 Zusammenfassung und Diskussion zu Forschungsfrage (2)

Die Forschungsfrage (2) legt den Fokus in der Lehrerkompetenzforschung auf die Dimensionalität und damit die Struktur des Konstrukts *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Experimentierkompetenzen.

(2) Inwiefern gibt es Hinweise auf Ein- oder Mehrdimensionalität des Konstrukts?

Joachim et al. (2020) fanden beim Vergleich einer ein- und dreidimensionalen Modellierung *von knowledge of what to assess* in Bezug auf Experimentierkompetenzen – unter Betrachtung von BIC und AIC – keine eindeutigen Ergebnisse in Bezug auf eine deutlich bessere Passung eines der beiden Modelle.

Die dreidimensionale IRT Modellierung des Konstrukts ergab latente Korrelation von 0.57 zwischen den Dimensionen Hypothesenbildung und Auswertung von Daten, 0.68 zwischen Hypothesenbildung und Planung von Experimenten und 0.80 zwischen Planung von Experimenten und Auswertung von Daten (vgl. Kapitel 6, Joachim et al., 2020). Die Werte deuten auf eine Mehrdimensionalität des Konstrukts hin. Eine Mehrdimensionalität könnte z.B. gegeben sein, wenn die Beurteilung der Kompetenzen in den einzelnen Phasen des Experimentierens spezifisches Wissen erfordert.

Vor diesem Hintergrund wurde die Dimensionalität von Experimentierkompetenzen betrachtet, die zu *knowledge of what to assess* in Bezug auf Experimentierkompetenzen beitragen (vgl. Kapitel 6, Joachim et al., 2020). Jedoch beziehen sich die vorliegenden Studien vornehmlich auf Schülerkompetenzen: Studien zum wissenschaftlichen Denken von Schüler*innen aus Jahrgang 5 bis 10 (Grube, 2010, S. 57) und Experimentieren von Schüler*innen aus Jahrgang 5 und 6 (Hammann et al., 2007, S. 42f.) gaben Hinweise auf Mehrdimensionalität der Konstrukte. Eine Untersuchung von Kompetenzen von Schüler*innen des 10. Jahrgangs ergab, dass ein vierdimensionales Modell, mit den Dimensionen Fragestellung, Hypothese, Untersuchungsdesign und Datenauswertung, keine bessere Passung zeigt, als ein eindimensionales Modell, das die vier Phasen der Erkenntnisgewinnung zur Dimension naturwissenschaftliche Untersuchungen zusammenfasst (Wellnitz, 2012, S. 130).

Mit zunehmendem Wissen zum Experimentieren in der Biologie ist ein umfassendes Verständnis der Methode der Erkenntnisgewinnung und Wissen über alle drei Phasen anzunehmen (vgl. Kapitel 6, Joachim et al., 2020). Entsprechend ist auch für Biologielehramtsstudierende, die in der Vermittlung und Beurteilung von Experimentierkompetenzen im Gesamten ausgebildet werden, zu erwarten, dass sie über Wissen für die Beurteilung aller drei Phasen des Experimentierens verfügen. Davon abweichen kann die Beurteilung spezifischer Kriterien unabhängig von den Phasen, die ggf. nicht (ausreichend) in der Lehrerbildung berücksichtigt werden (z.B. Begründung von Hypothesen).

Zusätzlich zu den theoretischen Überlegungen hat der sparsame, eindimensionale Ansatz Vorteile für die Testung sowie die Vermittlung von *knowledge of what to assess* in

Bezug auf Experimentierkompetenzen. Denn *knowledge of what to assess* im Rahmen von *assessment literacy* stellt insgesamt nur einen kleinen Teil des für angehende Biologielehrkräfte zu erwerbenden Wissens dar. Als eindimensionales Konstrukt kann *knowledge of what to assess* zudem mit akzeptabler Reliabilität forschungsökonomischer mit den anderen zu erwerbenden Wissensfacetten gemeinsam modelliert werden. Folglich wurde das eindimensionale Modell für die weiteren Analysen herangezogen (vgl. Kapitel 6, Joachim et al., 2020).

7.3 Zusammenfassung und Diskussion zu Forschungsfrage (3)

Die Forschungsfrage (3) widmet sich der Validität der im Rahmen der Arbeit vorgenommenen Interpretation der Testwerte zu *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Experimentierkompetenzen.

(3) Inwiefern gibt es Hinweise auf Validität?

Die in Kapitel 5 und 6 dargestellten Ergebnisse der Pilotstudie und Hauptstudie geben Hinweise auf die Validität der Interpretation der Testwerte. Die Validität wurde mittels eines Vergleichs bekannter Gruppen (Bachelor bzw. Staatsexamen ≤ Semester 6 und Master bzw. Staatsexamen ≥ Semester 7) und der Berechnung von Zusammenhängen des Konstrukts *knowledge of what to assess* mit verwandten Konstrukten wie *Diagnosekompetenz* und *Untersuchen* in der Biologie und Noten geprüft.

Die Befragung von sechs Expertinnen (drei Wissenschaftlerinnen, drei Lehrerinnen) zur Realitätsnähe und den Inhalten der Aufgaben lieferte vorab Hinweise für Inhaltsvalidität (vgl. Kapitel 4, Hasse et al., 2014).

In der Pilotstudie konnten in Teilen signifikant höhere Testleistungen für Masterstudierende im Vergleich zu Bachelorstudierenden bei der Bearbeitung von neun Testheften im *incomplete block design* (jedes Testheft enthielt neun der 27 Aufgaben der Pilotstudie, zweimal zwei Testhefte enthielten die gleichen Aufgaben jeweils in unterschiedlicher Reihenfolge) erzielt werden. So unterschieden sich die beiden Ausbildungsstufen bei der kombinierten Auswertung von zwei Testheften mit gleichen Aufgabenzusammenstellungen (Master: $M=0.58$, $SD=0.28$, $n=12$; Bachelor: $M=0.25$, $SD=0.17$, $n=14$; $p<0.001$). Keine signifikanten Unterschiede lagen hingegen für eine kombinierte Auswertung zweier weiterer Testhefte mit inhaltsgleichen Items vor (Master: $M=0.41$, $SD=0.17$, $n=9$; Bachelor: $M=0.35$, $SD=0.19$, $n=18$; n.s.) (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016).

In der Hauptstudie konnten höhere Fähigkeiten für Studierende auf Masterniveau (umfasst Studierende, die den Masterabschluss anstreben, bzw. Staatsexamen ab Semester 7, in Kapitel 6 als *students at the graduate level* bezeichnet) im Vergleich zu Studierenden auf Bachelorniveau (Studierende, die den Bachelorabschluss anstreben, bzw. Staatsexamen bis Semester 6, *students at the undergraduate level*) für die Stichprobe von n=495 festgestellt werden. Der Mann-Whitney-U-Test zeigte signifikant höhere Personenfähigkeiten für Studierende auf Masterniveau (n=224) (*graduate level*: MRank=276.07) als für Studierende auf Bachelorniveau (n=254) (*undergraduate level:* MRank=207.25), U=20255.50, Z=-5.447, p<0.001, r=0.25 (vgl. Kapitel 6, Joachim et al., 2020). Dieser Unterschied in den beiden Personengruppen stützt die Annahme höherer Personenfähigkeiten bezüglich *knowledge of what to assess* im Verlauf eines (konsekutiven) Studiums aufgrund von mehr entsprechenden Lerngelegenheiten. Der Erwerb von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen im Verlauf des Biologielehramtsstudiums kann als ein weiterer Indikator für die Validität der Interpretation von Testwerten angesehen werden (vgl. ebd.).

Die Zusammenhänge von *knowledge of what to assess* zu den Konstrukten *Diagnosekompetenz* und *Untersuchen* in der Biologie (Fragestellungen formulieren, Hypothesen generieren, Planen von Untersuchungen und Auswerten von Untersuchungen im Kontext von Biologie) konnte anhand der jeweiligen Testinhalte plausibel erklärt werden (vgl. Kapitel 6, Joachim et al., 2020). Ein Zusammenhang von *knowledge of what to assess* mit Selbstwirksamkeitserwartungen in Bezug zu Facetten der Unterrichtsplanung in Biologie (Mahler, 2014) für Studierende auf Masterniveau, im Gegensatz zu Studierenden auf Bachelorniveau, konnte durch eine bessere Selbsteinschätzung fortgeschrittener Studierender auf Masterniveau aufgrund von mehr Wissen zu und Erfahrungen in den adressierten Inhalten erklärt werden (vgl. Kapitel 6, Joachim et al., 2020).

Die Untersuchung von Zusammenhängen von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen mit schulischen und universitären *learning outcomes* (d.h. letzte Biologieschulnote, durchschnittliche Noten in Lehrveranstaltungen der Biologie und Didaktik der Biologie) zeigte den größten Zusammenhang von *knowledge of what to assess* mit der durchschnittlichen Note in Veranstaltungen der Didaktik der Biologie. Erklärungen für die Befunde werden in Kapitel 6 (Joachim et al., 2020) detaillierter ausgeführt (S. 92). Eine Erklärung liegt darin, dass das Testinstrument Inhalte von Lehrveranstaltungen in Transferkontexten aufgreift.

Insgesamt liefern die vorgenommenen Untersuchungen multiple Hinweise auf Validität.

7.4 Zusammenfassung und Diskussion zu Forschungsfrage (4)

Für die Ausbildung von Lehrkräften ist es relevant zu erfahren, inwiefern angehende Biologielehrkräfte über *knowledge of what to assess* in Bezug auf Experimentierkompetenzen verfügen. Mit Forschungsfrage (4) wird die Qualität von *knowledge of what to assess* von Biologielehramtsstudierenden untersucht.

(4) Welche Stärken und Schwächen haben Biologielehramtsstudierende in der Beurteilung von Experimentierkompetenzen von Schüler*innen?

In den Analysen der Hauptstudie (vgl. Kapitel 6, Joachim et al., 2020) konnten für vier Kategorien von Items Schwierigkeiten beschrieben werden. Den Studierenden fiel in der Beurteilung die Kategorie iv) beurteilen und korrigieren fehlerhafter Schüler*innenlösungen am leichtesten. Signifikant schwieriger waren die Kategorien iii) beurteilen der Planung in Bezug auf Standardisierung und Genauigkeit, ii) beurteilen korrekter Schüler*innenlösungen und i) beurteilen von Vorstellungen von Schüler*innen. Im Gegensatz zu den Kategorien unterscheiden sich die drei Phasen des Experimentierens in der Schwierigkeit der abzugebenden Beurteilungen von Schülerexperimentierkompetenzen nicht signifikant. Auffällig war die geringe Verwendung von Fachbegriffen wie z.B. Ingenieursmodus, Confirmation Bias, Fehlerdiskussion und Kontrollansatz in den Beurteilungen. Fachbegriffe für Vorstellungen von Schüler*innen und Kriterien des Experimentierens sind jedoch ausbildungsrelevant für präzise, fachangemessene Beurteilungen und sollten Teil von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen sein. Viele Biologielehramtsstudierende beschrieben oder erklärten die Vorstellungen und Kriterien des Experimentierens, benannten sie aber nicht. Beispielsweise nahmen 456 Studierende in ihrer Beurteilung der Schülerkompetenzen Bezug zur unsystematischen Variation von Variablen. Knapp ein Viertel der Personen verwendete in dem Zusammenhang einen der Begriffe Kontrollansatz bzw. Kontrollgruppe (Joachim et al., 2020, S. 15).

Die unterschiedlichen Schwierigkeiten von Items könnten mit ihren Inhalten und Anforderungen erklärt werden. Wissen zu Vorstellungen von Schüler*innen zum Experimentieren wird vornehmlich in Veranstaltungen der Didaktik der Biologie erworben, die lediglich einen kleinen Anteil des Studiums ausmachen (Kultusministerkonferenz, 2019b). Wissen zu Kriterien des Experimentierens kann sowohl in Veranstaltungen der Didaktik der Biologie als auch fachwissenschaftlichen Veranstaltungen erworben werden und teilweise bereits aus dem Biologieunterricht in der Schule vorliegen (vgl. Kapitel 6, Joachim et al., 2020). Die Fähigkeit von Schüler*innen beispielsweise zur systematischen Varia-

tion von Variablen (Item 7 der Hauptstudie) wurde vielfach untersucht und wird als wichtiges Kriterium des Experimentierens in für die Lehrerbildung relevanten Veröffentlichungen betont (Chen & Klahr, 1999; Schulz et al., 2012). Entsprechend kann die geringe Schwierigkeit der Beurteilung der unsystematischen Variation von Variablen mit erfolgreicher Implementation in der Lehrerbildung erklärt werden. Die Anwendung der experimentspezifischen Kriterien *Begründung* und *Testbarkeit* von Hypothesen hingegen ist bislang weniger untersucht. Zudem sind diese Kriterien bzw. Anforderungen an die experimentelle Vorgehensweise nicht eindeutig in Lehrwerken dargestellt (vgl. Baack & Steinert, 2015). Der Fokus auf Schülerfehler in den Beurteilungen der Lehramtsstudierenden kann ggf. auf eigene Erfahrungen in der Schulzeit zurückgeführt werden. Der Zweck von Beurteilungen von Schüler*innen wird oft in der Notengebung gesehen anstatt der Generierung von Informationen über den Lernstand der Schüler*innen für adaptive Fördermaßnahmen seitens der Lehrkraft bzw. anstatt als Ausgangspunkt für passgenaueres Lernen (Xu & Brown, 2016, S. 154).

Ein weiterer Faktor, der die Itemschwierigkeit beeinflussen kann, sind die gestellten Herausforderungen in den Beurteilungsaufgaben. Einige Beurteilungsaufgaben erfordern die Berücksichtigung mehrerer Aspekte. Von den Studierenden wird z.B. eine umfassende Beurteilung unter Berücksichtigung mehrerer Fehler oder von Fehlern und korrekten Lösungen von Schüler*innen erwartet. Eine Konzentration der Studierenden dabei auf einen Aspekt – den für sie eindeutigsten bzw. zuerst wahrgenommenen – kann zu höheren Schwierigkeiten in der Beurteilung von zusätzlichen Fehlern bzw. korrekten Lösungen der Schüler*innen geführt haben (vgl. Kapitel 6, Joachim et al., 2020).

Mit den Erkenntnissen zu Stärken und Schwächen der Biologielehramtsstudierenden in der Beurteilung von Experimentierkompetenzen liefert die Hauptstudie einen Beitrag für die Ausbildung von Lehrkräften. Der Fokus liegt auf Statusdiagnosen. Aussagen z.B. über prozessdiagnostische Kompetenzen oder die Beurteilung von Lernfortschritten von Schüler*innen beim Experimentieren können nicht getroffen werden.

8. Fazit und Ausblick

Das Ziel dieser Arbeit war die Modellierung, Messung und Validierung von *knowledge of what to assess* von Biologielehramtsstudierenden in Bezug auf Experimentierkompetenzen. Zunächst wurden Aufgaben zur Erfassung des Konstrukts entwickelt. Die Pilotstudie und Hauptstudie geben empirische Hinweise auf die Eignung des Instruments zur Erfassung von Wissen und Fähigkeiten von Biologielehramtsstudierenden für die Beurteilung von Experimentierkompetenzen von Schüler*innen. Erkenntnisse zu Stärken und Schwächen von Biologielehramtsstudierenden in der Beurteilung von Schülerexperimentierkompetenzen liegen für die Förderung vor.

Für die Ausbildung von Lehrkräften lässt sich ableiten, dass die Vermittlung von Schülervorstellungen zum Experimentieren weiter vertieft werden kann und ein Augenmerk auf die Verwendung von Fachbegriffen gelegt werden sollte, um eine solide Wissensbasis für die Beurteilung der Experimentierkompetenzen zu schaffen. Diese ist eine zentrale Grundlage für die Förderung von experimentellen Kompetenzen bei Schüler*innen (vgl. Kapitel 6, Joachim et al., 2020).

Aus der Pilotstudie liegt ein Pool an getesteten Aufgaben vor. Diese können für die Erfassung oder Vermittlung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen angepasst und verwendet werden. Sie können für Interventions- oder Längsschnittstudien weiterentwickelt werden, um die Entwicklung von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen, mit dem Ziel der Förderung einer wichtigen Komponente von *assessment literacy* in der universitären Ausbildung von Lehrkräften, zu untersuchen (vgl. Kapitel 5, Bögeholz, Joachim et al., 2016).

Zukünftige Analysen könnten zudem unterschiedliche Studiengänge für Lehrkräfte verschiedener Schulformen oder die Entwicklung von *knowledge of what to assess* mit zunehmender Praxiserfahrung in den Blick nehmen.

Die vorliegende Arbeit untersucht einen Bereich der Wissensbasis von *assessment literacy* (vgl. Abell & Siegel, 2011; Xu & Brown, 2016). Weiterer Forschungsbedarf besteht in Bezug auf die anderen Bereiche von *assessment literacy* sowie in Bezug auf den Zusammenhang von Beurteilungen mit anderen Fähigkeiten. Für den Lernerfolg von Schüler*innen ist entscheidend, dass Lehrkräfte die in Beurteilungen gewonnenen Erkenntnisse für die Adaptation des Unterrichts nutzen können (vgl. *knowledge of assessment interpretation and action-taking,* Abell & Siegel, 2011; Bromme, 1997, S. 200). Nur Lehrkräfte, die sowohl über Fähigkeiten im Beurteilen als auch im Unterrichten verfügen,

können ihre Schüler*innen optimal fördern (Weinert et al., 1990). Entsprechend relevant ist die Untersuchung des Zusammenhangs von Wissen und Fähigkeiten für die Beurteilung und die Vermittlung von Experimentierkompetenzen. Vor diesem Hintergrund stehen im Rahmen des ExMo Projekts noch gemeinsame Modellierungen von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen mit den postulierten Dimensionen Analyse und Planung von experimentellem Biologieunterricht (Hasse et al., 2014) aus (vgl. Kapitel 6, Joachim et al., 2020). Bislang ungeklärt ist u.a., inwiefern Analyse und Planung als Vermittlungskompetenzen in einer Dimension zusammenfallen und inwiefern dies von *knowledge of what to assess* in Bezug auf Experimentierkompetenzen abgrenzbar ist. *Knowledge of what to assess* ist kontextgebunden (z.B. an den Kontext Experimentieren). Eine eindimensionale Modellierung von *knowledge of what to assess* mit Vermittlungskompetenzen würde für eine zentrale Bedeutung von Wissen zum Experimentieren als Grundlage sowohl von *knowledge of what to assess* als auch Vermittlungskompetenzen sprechen.

Neben Wissen stellen auch Überzeugungen (vgl. Werte und Prinzipien bei Abell & Siegel, 2011) einen bedeutenden Aspekt professioneller Kompetenz dar: In einer Untersuchung professioneller Kompetenz frühpädagogischer Fachkräfte im Bereich Mathematik wurde die Relevanz von Wissen und Überzeugungen für die Situationswahrnehmung und Handlungsplanung herausgestellt (Dunekacke, 2015). Vergleichbare Analysen zu Überzeugungen in Bezug auf Experimentierkompetenzen im Bereich Biologie könnten weiteren Aufschluss über Faktoren geben, die zur Förderung von Experimentierkompetenzen bei Schüler*innen beitragen.

Abschließend kann festgehalten werden, dass die Arbeit, unter Einsatz eines Instruments mit realitätsnahen Aufgaben in drei unterrichtsrelevanten Kontexten der Biologie, Erkenntnisse zur Dimensionalität von *knowledge of what to assess* sowie zentrales Wissen für die Ausgestaltung der Ausbildung von Biologielehrkräften für verschiedene Schulformen bundesweit generieren konnte. Befunde zu Stärken und Schwächen hinsichtlich *knowledge of what to assess* in Bezug auf Experimentierkompetenzen können nach entsprechender Aufbereitung für die Lehrerbildung genutzt werden.

Literaturverzeichnis

Abell, S. K. (2007). Research on science teacher knowledge. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 1105-1149). Lawrence Erlbaum.

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Springer.

Aufschnaiter, C. v., Cappell, J., Dübbelde, G., Ennemoser, M., Mayer, J., Stiensmeier-Pelster, J., Sträßer, R., & Wolgast, A. (2015). Diagnostische Kompetenz: Theoretische Überlegungen zu einem zentralen Konstrukt der Lehrerbildung. *Zeitschrift für Pädagogik, 61*(5), 738-758.

Baack, K., & Steinert, K. (2015). *Natura 7/8 Biologie für Gymnasien, Niedersachsen.* Klett.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469-520.

Beck, E., Baer, M., Guldimann, T., Bischoff, S., Brühwiler, C., Müller, P., Niedermann, R., Rogalla, M., & Vogt, F. (2008). *Adaptive Lehrkompetenz: Analyse und Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens.* Waxmann.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Open University Press.

Bögeholz, S., Carstensen, C., Hammann, M., Hasse, S., & Joachim, C. (2013). ExMo – Teaching Competencies and Assessment Competencies in Experimental Biology Lessons: Modeling, Validation and Development of a Test Instrument. In S. Blömeke & O. Zlatkin-Troitschanskaia (Eds.), *The German funding initiative "Modeling and measuring competencies in higher education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students* (pp. 43-46). Humboldt-Universität zu Berlin; Johannes Gutenberg-Universität Mainz.

Bögeholz*, S., Joachim*, C., Hasse, S., & Hammann, M. (2016). Kompetenzen von (angehenden) Biologielehrkräften zur Beurteilung von Experimentierkompetenzen. *Unterrichtswissenschaft, 44*(1), 40-54. (*gemeinsame Erstautorenschaft)

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Springer.

Bouffard-Bouchard, T., Parent, S., & Larivee, S. (1991). Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. *International Journal of Behavioral Development, 14*(2), 153-164.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (S. 177-212). Hogrefe.

Bruckermann, T., Arnold, J., Kremer, K., & Schlüter, K. (2017). Forschendes Lernen in der Biologie. In T. Bruckermann & K. Schlüter (Hrsg.), *Forschendes Lernen im Experimentalpraktikum Biologie* (S. 11-26). Springer.

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften - Ergebnisse des Forschungsprogramms COACTIV* (S. 215-234). Waxmann.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte Aufl.). Pearson.

Cappell, J. (2013). Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 146). Logos.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49.

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*(6), 623-654.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005). *How students learn: History, mathematics, and science in the classroom*. National Academies Press.

Dübbelde, G. (2013). *Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung* [Dissertation, Universität Kassel]. KOBRA. https://kobra.uni-kassel.de/handle/123456789/2013122044701

Dübbelde, G., Mayer, J., Möller, A., & Aufschnaiter, C. v. (2010). Diagnosekompetenz von Biologie-Lehramtsstudierenden zum Kompetenzbereich Erkenntnisgewinnung. In D. Krüger, A. Upmeier zu Belzen & S. Nitz (Hrsg.), *Erkenntnisweg Biologiedidaktik 9* (S. 119-134). Universitätsdruckerei Kassel.

Duggan, S., & Gott, R. (2000). Intermediate general national vocational qualification (GNVQ) science: A missed opportunity for a focus on procedural understanding? *Research in Science & Technological Education, 18*(2), 201-214.

Dunekacke, S. (2015). *Mathematische Bildung in Alltags- und Spielsituationen begleiten – Handlungsnahe Erfassung mathematikdidaktischer Kompetenz angehender frühpädagogischer Fachkräfte durch die Bearbeitung von Videovignetten* [Dissertation, Humboldt-Universität zu Berlin]. edoc. https://edoc.hu-berlin.de/bitstream/handle/18452/18232/dunekacke.pdf?sequence=1

Emden, M. (2011). Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens: Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 118). Logos.

Fischer, C., Kopmann, H., Rott, D., Veber, M., & Zeinz, H. (2014). Adaptive Lehrkompetenz und pädagogische Haltung. In K. Zierer (Hrsg.), *Jahrbuch für allgemeine Didaktik 2014. Thementeil Allgemeine Didaktik für eine inklusive Schule* (S. 16-34). Schneider Verlag Hohengehren.

Großschedl, F., Harms, U., Kleickmann, T., & Glowinski, I. (2015). Preservice biology teachers' professional knowledge: Structure and learning opportunities. *Journal of Science Teacher Education, 26*(3), 291-318. https://doi.org/10.1007/s10972-015-9423-6

Grube, C. R. (2010). *Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung. Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I*. [Dissertation, Universität Kassel]. KOBRA. https://kobra.uni-kassel.de/handle/123456789/2011041537247

Hammann, M. (2004). Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung – dargestellt anhand von Kompetenzen beim Experimentieren. *MNU, 57*(4), 196-203.

Hammann, M. (2007). Das Scientific Discovery as Dual Search-Modell. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 187-196). Springer.

Hammann, M., & Mayer, J. (2012). Was lernen Schülerinnen und Schüler beim Experimentieren? *Biologie in unserer Zeit*, *42*(5), 284-285.

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

Hammann, M., Phan, T.T.H., Ehmer, M., & Bayrhuber, H. (2006). Fehlerfrei Experimentieren. *MNU, 59*(5), 292-299.

Hasse, S., Joachim, C., Bögeholz, S., & Hammann, M. (2014). Assessing teaching and assessment competences of biology teacher trainees: Lessons from item development. *International Journal of Education in Mathematics, Science and Technology, 2*(3), 191-205.

Joachim, C., Hammann, M., Carstensen, C. H., & Bögeholz, S. (2020). Modeling and measuring pre-service teachers' assessment literacy regarding experimentation competences in biology. *Education Sciences, 10*(5). https://doi.org/10.3390/educsci10050140

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. The MIT Press.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for Scientific Experimentation: A Developmental Study. *Cognitive Psychology, 25*, 111-146.

Klautke, S. (1997). Ist das Experimentieren im Biologieunterricht noch zeitgemäß? *MNU*, *50*(6)*,* 323-329.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2004). *The development of National Educational Standards: An expertise.* Bundesministerium für Bildung und Forschung.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. The MIT Press.

Krüger, D. (2007). Die Conceptual Change-Theorie. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 81-92). Springer.

Krüger, D. (2009). Bezaubernde Biologie - mit Hypothesen der Lösung auf der Spur. *MNU, 62*(1), 41-46.

Krüger, D., Upmeier zu Belzen, A., Nordmeier, V., Tiemann, R., Hartmann, S., Mathesius, S., Stiller, J., & Straube, P. (2014). *Kooperation der Projekte Ko-WADiS und ExMo*. Unveröffentlicht.

Kultusministerkonferenz (Hrsg.). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. Wolters Kluwer Deutschland GmbH. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

Kultusministerkonferenz (Hrsg.). (2019a). *Standards für die Lehrerbildung: Bildungswissenschaften*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf

Kultusministerkonferenz (Hrsg.). (2019b). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung.* https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf

Kunter, M., Kleickmann, T., Klusmann, U., & Richter, D. (2011). Die Entwicklung professioneller Kompetenz von Lehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 55-68). Waxmann.

Li, J., & Klahr, D. (2006). The psychology of scientific thinking: Implications for science teaching and learning. In J. Rhoton & P. Shane (Eds.), *Teaching science in the 21st century* (pp.307-328). National Science Teachers Association and National Science Education Leadership Association: NSTA Press.

List, M. K. (2018). *Testbearbeitungsverhalten in Leistungstests: Modellierung von Testabbruch und Leistungsabfall* [Dissertation, Christian-Albrechts-Universität zu Kiel]. MACAU. https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00007735/diss_mk_list_testbearbeitungsverhalten_in_leistungstests.pdf

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

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95-132). Kluwer.

Mahler, H. (2014). *Selbstwirksamkeitserwartungen angehender Biologielehrkräfte—Entwicklung eines Messinstrumentes*. [Masterarbeit, Georg-August-Universität Göttingen]. Unveröffentlicht.

Marbach-Ad, G., & Claassen, L. A. (2001). Improving students' questions in inquiry labs. *The American Biology Teacher, 63*(6), 410-419.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden* (S. 177-186). Springer.

Mayer, J., Grube, C., & Möller, A. (2008). Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In U. Harms & A. Sandmann (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 3. Ausbildung und Professionalisierung von Lehrkräften* (S. 63-79). StudienVerlag.

Mayer, J., & Ziemek, H.-P. (2006). Offenes Experimentieren: Forschendes Lernen im Biologieunterricht. *Unterricht Biologie, 317*, 4-12.

Méhaut, P., & Winch, C. (2012). The European qualification framework: Skills, competences or knowledge? *European Educational Research Journal*, *11*(3), 369-381. https://doi.org/10.2304/eerj.2012.11.3.369

National Research Council. (1996). *National Science Education Standards*. National Academy Press.

Niedersächsisches Kultusministerium (Hrsg.). (2007). *Kerncurriculum für das Gymnasium Schuljahrgänge 5 -10: Naturwissenschaften*. http://db2.nibis.de/1db/cuvo/datei/kc_gym_nws_07_nib.pdf

OECD. (2005a). *The definition and selection of key competencies: Executive summary*. http://www.oecd.org/pisa/35070367.pdf

OECD. (2005b). *Definition und Auswahl von Schlüsselkompetenzen: Zusammenfassung*. http://www.oecd.org/pisa/35693281.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Praetorius, A.-K., & Südkamp, A. (2017). Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (S. 13-18). Waxmann.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84. https://doi.org/10.1002/tea.20163

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*(9), 859-882.

Schrader, F.-W. (2006). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3. überarbeitete und erweiterte Aufl., S. 95-100). Beltz.

Schrader, F.-W. (2008). Diagnoseleistungen und diagnostische Kompetenzen von Lehrkräften. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 168-177). Hogrefe.

Schrader, F.-W. (2009). Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*, 237-245.

Schrader, F.-W. (2011). Lehrer als Diagnostiker. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 683-698). Waxmann.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerinnen- und Lehrerbildung, 31*(2), 154-165.

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik, 1*(1), 27-52.

Schrader, F.-W., & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Urteile. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 22*(4), 312-324.

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

Shavelson, R. F. (2009). *Measuring College Learning Responsibly: Accountability in a New Era*. Stanford University Press.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz*. Zeitschrift für Pädagogische Psychologie, 19*, 85-95.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*, 1-14.

Straube, P. F. (2016). Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-)Studierenden im Fach Physik. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 209). Logos.

Töpperwien, B., & Köttker, N. (2010). *Kompetenzen vermitteln, Kompetenzen erwerben: Biologie*. Aulis.

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202-248.

Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht: Struktur und Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften, 16,* 41-57.

Voss, T. (2019). Lehrkraftwissen und dessen Erwerb. In N. McElvany, F. Schwabe, W. Bos & H. G. Holtappels (Hrsg.), *Lehrerbildung – Potentiale und Herausforderungen in den drei Phasen* (S. 9-28). Waxmann.

Weinert, F. E. (2000). Lehren und Lernen für die Zukunft - Ansprüche an das Lernen in der Schule. *Pädagogische Nachrichten Rheinland-Pfalz, Heft 2 - Schulleben Schulkultur*, Sonderseiten 1-16.

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1990). Educational expertise: Closing the gap between educational research and classroom practice. *School Psychology International, 11*, 163-180.

Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung* (Biologie lernen und lehren, Bd. 2). Logos.

Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A., Sumfleth, E., & Walpuski, M. (2012). Evaluation der Bildungsstandards - eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften, 18*, 261-291.

Wember, F. B., & Melle, I. (2018). Adaptive Lernsituationen im inklusiven Unterricht: Planung und Analyse von Unterricht auf Basis des Universal Design for Learning. In S. Hußmann & B. Welzel (Hrsg.), *DoProfiL – Das Dortmunder Profil für inklusionsorientierte Lehrerinnen- und Lehrerbildung* (S. 57-72). Waxmann.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162.

Danksagung

Ich möchte mich herzlich bei allen Personen bedanken, die zur Entstehung dieser Arbeit beigetragen haben.

Besonderer Dank gilt meiner Betreuerin Prof. Dr. Susanne Bögeholz für die Unterstützung und Beratung in allen Phasen der Arbeit sowie meinem Zweitbetreuer Prof. Dr. Marcus Hammann und den weiteren Mitgliedern des ExMo Projekts, Prof. Dr. Claus H. Carstensen und Sascha Hasse, für die konstruktiven Gespräche, u.a. in Bezug auf die Konzeption der Aufgaben und das methodische Vorgehen. Vielen Dank für die tolle und lehrreiche Zeit, auf die ich gerne zurückblicke.

An den empirischen Studien haben über 600 Studierende teilgenommen. Ihnen und den Mitarbeiter*innen der Universitäten und Hochschulen, die die Befragungen ermöglicht haben, möchte ich hiermit danken.

In diesem Zusammenhang danke ich auch allen am ExMo Projekt beteiligten Hilfskräften für die großartige Hilfe beim Scoren der Antworten und meinen Kolleg*innen aus der Abteilung der Didaktik der Biologie für ihre Unterstützung.

Schließlich geht ein großes Dankeschön an meine Familie, die immer für mich da ist.