Überprüfung der Vorhersagekraft eines Kompetenztests zur Performanz beim Experimentieren

René Mückai & Dirk Krüger

rene.mueckai@fu-berlin.de

Freie Universität Berlin, Didaktik der Biologie,

Schwendenerstraße 1, 14195 Berlin

Zusammenfassung

*Im Projekt ValiDiS wird untersucht, inwieweit ein im Vorgängerprojekt Ko-WADiS entwickeltes schriftliches Testinstrument zur Erfassung der Kompetenz im Umgang mit naturwissenschaftlichen Denk- und Arbeitsweisen das problemlösende Verhalten von Studierenden in einer realen Experimentiersituation voraussagen kann (prognostische Validität). Zur Beurteilung der individuellen Performanz werden N = 30 Biologie-Lehramtsstudierende in einer Experimentiersituation video- und audiodokumentiert und hinsichtlich lösungsrelevanter kognitiver Fähigkeiten und sensomotorischer Fertigkeiten kategoriengeleitet analysiert. Basierend auf ihren jeweiligen Stärken lassen sich zwei Experimentiertypen identifizieren: Praktiker*innen (n = 18) und Theoretiker*innen (n = 12). Korrelationsanalysen zeigen, dass zwischen den Ko-WADiS-Testwerten und der Performanz in der praktischen Experimentiersituation keine oder nur kleine Zusammenhänge bestehen. Eine Vorhersage des problemlösenden Verhaltens durch den Ko-WADiS-Test ist nicht möglich.*

Abstract

The project ValiDiS is investigating the extent to which a written test instrument developed in the predecessor project Ko-WADiS to measure scientific reasoning competencies can predict the problem-solving behaviour of university students in a real experimental situation (prognostic validity). In order to assess individual performance, N = 30 biology pre-service teachers are video- and audiotaped in an experimental situation. Their performance is analyzed in terms of relevant cognitive abilities and psychomotor skills in a category-based manner. Based on their respective strengths, two types of experimenters can be identified: practitioners (n = 18) and theorists (n = 12). Correlation analyses show that there are no or only small correlations between the Ko-WADiS Test-scores and the performance in the experimental situation. The Ko-WADiS test cannot predict the problem-solving behaviour.

1 Einleitung

Das hypothesengeleitete Experimentieren zählt neben dem Beobachten, Vergleichen, Ordnen und Modellieren zu den wesentlichen Arbeitsweisen der Biologie (UPMEIER ZU BELZEN & KRÜGER, 2019; WELLNITZ & MAYER, 2016). Die Entwicklung von Kompetenzen hinsichtlich der Planung, Durchführung und Auswertung dieser Arbeitsweisen ist ein essentieller Bestandteil einer naturwissenschaftlichen Grundbildung und somit auch des Biologieunterrichts (BYBEE, 2002; KMK, 2005).

Als Teil ihrer universitären Ausbildung soll es Lehramtsstudierenden der Biologie ermöglicht werden, entsprechende Kompetenzen zu entwickeln, um Schülerînnen angemessen unterrichten sowie deren Kompetenzentwicklung diagnostizieren und fördern zu können (BAUMERT & KUNTER, 2006; KMK, 2019). Während bereits zahlreiche Projekte zur Modellierung und Erfassung der Kompetenzen von Schülerînnen im Bereich der Erkenntnisgewinnung durchgeführt wurden (z.B. HAMMANN et al., 2008; SCHMIDT, 2016; WELLNITZ & MAYER, 2013), ist die Zahl der Studien, die sich der Kompetenzentwicklung von (Lehramts-)Studierenden widmen, überschaubar (vgl. HARTMANN et al., 2015; KRELL et al., 2018).

Im Projekt ValiDiS[1] wird die Validität der Testwertinterpretationen eines im Vorgängerprojekt Ko-WADiS entwickelten schriftlichen Instruments zur Erfassung der Kompetenz von Studierenden im Umgang mit naturwissenschaftlichen Denk- und Arbeitsweisen vertieft untersucht. In ValiDiS wird der in Ko-WADiS begonnene Längsschnitt fortgeführt, um die Sensitivität des Instruments für theoriebasiert vorhergesagte Gruppenunterschiede und die Bedeutung von Kovariaten bei der Kompetenzentwicklung zu untersuchen (KRÜGER et al., 2020). Ebenso wird untersucht, inwieweit die Interpretation der Testwerte auf die Effektivität von experimentell angelegten Interventionen zur Förderung von Kompetenzen des wissenschaftlichen Denkens hinweisen kann.

Das Desiderat dieser Studie ist es, zu prüfen, ob und inwieweit die erzielten Testwerte im Ko-WADiS-Test die Performanz von Lehramtsstudierenden in einer realen Experimentiersituation vorhersagen können (*prognostische Validität,* vgl. SCHMIEMANN & LÜCKEN, 2014).

[1] Das Akronym ValiDiS steht für: *Vali*dierungsstudie zum wissenschaftlichen *D*enken *i*m naturwissenschaftlichen *S*tudium. Die Autoren danken dem Bundesministerium für Bildung und Forschung für die finanzielle Förderung des ValiDiS-Projektes (Förderkennzeichen: 01PK15004A) im Rahmen des Forschungsprogramms KoKoHs.

2 Theorie

„There is a general pattern to all scientific reasoning." (GIERE ET AL., 2006, S. 6). Dieses allen naturwissenschaftlichen Arbeitsweisen zugrundeliegende Muster kann als kontext-spezifische Form des Problemlösens beschrieben werden (HARTMANN et al., 2015). Im Folgenden sollen die Charakteristika und Möglichkeiten der Erfassung des experimentellen Problemlösens von Lehramtsstudierenden genauer beleuchtet werden.

2.1 Experimentieren als problemlösendes Verhalten

Zur Beschreibung eines experimentell-problemlösenden Verhaltens wird das Kompetenzstrukturmodell zum wissenschaftlichen Denken von MAYER (2007) zugrunde gelegt. Demnach erfordern naturwissenschaftliche Arbeitsweisen im Allgemeinen und Experimentieren im Speziellen spezielle kognitive Fähigkeiten (*inquiry skills*), die sich in den vier Teilkompetenzen „Naturwissenschaftliche Fragen formulieren", „Hypothesen generieren", „Untersuchungen planen" und „Daten analysieren/ Schlussfolgerungen ziehen" widerspiegeln.

Um ein reales Problem mithilfe eines Experimentes lösen zu können, werden neben den beschriebenen kognitiven Fähigkeiten auch sensomotorische Fertigkeiten (*manual skills*) benötigt. Diese werden durch die Einführung einer fünften Teilkompetenz „Untersuchungen durchführen" in das Kompetenzstrukturmodell ergänzt (vgl. MEIER & MAYER, 2012; SCHMIDT, 2016).

2.2 Erfassung von experimentell-problemlösendem Verhalten

Generell lassen sich zwei Aufgabenformate zur Erfassung von experimentell-problemlösendem Verhalten unterscheiden: paper-pencil-Tests und praktische Experimentiertests (Performance Assessment, vgl. HEIDRICH, 2017).

Auf paper-pencil-Tests basierende Kompetenzmodelle können durch ihre Fokussierung auf die kognitiven Fähigkeiten den Umgang der Lernenden mit naturwissenschaftlichen Arbeitsweisen nicht vollumfänglich abbilden. Die motivationalen, praktisch-orientierten und sensomotorischen Fertigkeiten werden in diesen Modellen häufig nicht berücksichtigt, da sie schwer in schriftlichen Testformaten erfasst werden können (SCHECKER & PARCHMANN, 2006).

Praktische Experimentiertests im Sinne des Performance Assessments grenzen sich gegenüber paper-pencil-Tests durch ein authentisches, offenes und anwendungsbezogenes Aufgabendesign mit mehreren Antwortmöglichkeiten ab (SOLANO-FLORES & SHAVELSON, 1997). Dass es sich hierbei um eine differente Leistung von Lernenden gegenüber schriftlich erfassbaren, naturwissenschaftlichen Fähigkeiten handelt, konnte in mehreren Studien über

geringe Korrelationen zwischen den verschiedenen Testformaten bestätigt werden (z.B. HAMMANN et al., 2008; LAWRENCE et al., 2001).

2.3 Vorhersagekraft von Testwertinterpretationen

Inwieweit die Ergebnisse von schriftlichen Kompetenztests zum Experimentieren mit der Leistung in einem Außenkriterium (z.B.: problemlösendes Verhalten) zusammenhängt, ist eine grundlegende Frage der Kriteriumsvalidität (AERA et al., 2014). Es werden zwei Formen der Kriteriumsvalidität unterschieden: die prognostische Validität (predictive validity), die sich darin bemisst, ob Testwerten das spätere Verhalten im Außenkriterium korrekt vorhersagen, und die Übereinstimmungsvalidität (concurrent validity), bei der Test-Score und Kriteriumswert zum selben Messzeitpunkt erhoben werden (BORTZ & DÖRING, 2006). In dieser Untersuchung geht es um die prognostische Validität der Ko-WADiS-Testwerte.

3 Fragestellungen und Hypothesen

Der vorliegenden Studie liegt die folgende Fragestellung zugrunde:

F_1 Inwieweit ist das problemlösende Verhalten der Studierenden beim praktischen Experimentieren durch ihre Testwerte im Ko-WADiS-Test vorhersagbar?

HAMMANN et al. (2008) und LAWRENCE et al. (2001) konnten zeigen, dass für das experimentell-problemlösende Verhalten von Schülerînnen nur ein geringer Zusammenhang zwischen den Testwerten in einem paper-pencil-MC-Test und der Performanz in einem praktischen Experimentiertest besteht. Daher werden kleine Korrelationen ($0.1 < \rho < 0.3$) zwischen dem problemlösenden Verhalten der Studierenden in einer realen Experimentiersituation und den Ko-WADiS-Testwerten erwartet.

Das dem Ko-WADiS-Test zugrundeliegende Kompetenzmodell (vgl. HARTMANN et al., 2015) umfasst nur kognitive Fähigkeiten, die sensomotorische Teilkompetenz „Untersuchungen durchführen“ wird nicht berücksichtigt. Praktische Experimentiertests im Sinne des Performance Assessment setzen die Anwendung sowohl von Fähigkeiten als auch von Fertigkeiten voraus. Da es sich dabei um unterschiedliche Leistungen der Studierenden handelt, lässt sich daraus die Frage ableiten:

F_2 Inwieweit lassen sich in einem praktischen Experimentiertest voneinander abgrenzbare Typen identifizieren?

4 Methodik

4.1 Stichprobe, Setting und Datenerhebung

Zur quantitativen Erfassung der Kompetenzen im Bereich Experimentieren wird der im Vorgängerprojekt entwickelte Ko-WADiS-Test eingesetzt (HARTMANN et al., 2015b). Dieser besteht aus 21 Aufgaben zum wissenschaftlichen Denken im Multiple-Choice-Format.

Aus dem in ValiDiS fortgeführten Längsschnitt mit über 500 Studierenden wurden über einen Zeitraum von fünf Semestern 30 Biologie-Lehramtsstudierende entsprechend eines *maximum variation sampling* ausgewählt, um an der Validierungsstudie teilzunehmen (SURI, 2011). Basierend auf den Ergebnissen des Ko-WADiS-Tests wurden Proband*innen eingeladen, deren Gesamtheit die volle Breite des Leistungsspektrums abbildet. Um zu prüfen, inwieweit die Proband*innen selbstständig ein naturwissenschaftliches Experiment planen, durchführen und auswerten können, wird ein Experimentiersetting entsprechend den Vorgaben des Performance Assessment nach SOLANO-FLORES UND SHAVELSON (1997) entworfen. Demnach müssen drei Anforderungen an ein Performance Assessment Setting erfüllt sein: a) eine kontextualisierte Problemstellung, deren Lösung die Verwendung konkreter (vorgegebener) Materialien erfordert, b) ein Antwortformat, in dem der Lösungsprozess und die Antworten der Proband*innen erfasst werden und c) ein Bewertungssystem, um die wissenschaftliche Qualität des jeweiligen Lösungsprozesses zu bewerten.

Bei dem in dieser Studie verwendeten Experimentiersetting handelt sich um eine Problemstellung mit biologischem Kontext, in der die Reaktionszeit des Menschen durch Fangen eines Lineals unter Einfluss von Traubenzucker und/oder Koffein untersucht wird (MÖLLER & SPECHT, 2013). Die Proband*innen sind angehalten, ihr Experiment möglichst wissenschaftlich durchzuführen und ihr Vorgehen angemessen zu dokumentieren. Die Proband*innen haben zur Aufgabenbearbeitung eine Auswahl an Materialien und eine Bearbeitungszeit von 60 Minuten zur Verfügung.

Zur Erhebung des experimentell-problemlösenden Verhaltens der Studierenden wird als Antwortformat eine passiv-teilnehmende, vermittelnde Beobachtung mit gleichzeitigem lauten Denken unter Einsatz von technischen Hilfsmitteln (Kamera, Mikrofon) gewählt (ROTH & HOLLING, 1999). Die sprachlichen Äußerungen werden wortwörtlich unter Anleitung eines detaillierten Manuals mit dem Programm f4transkript (DR. DRESING & PEHL GMBH, 2018) transkribiert.

4.2 Datenauswertung

Beide Datenquellen, die Videografien und die Transkripte, werden mit Hilfe der Software MAXQDA (VERBI SOFTWARE, 2018) im Sinne der inhaltlich strukturierenden qualitativen Inhaltsanalyse (KUCKARTZ, 2014) unter Anwendung eines deduktiv-induktiv entwickelten Kodierleitfadens ausgewertet. Die aus der Literatur (z.B. ARNOLD et al., 2013; MAYER, 2007) und einer Pilotierung im Sommersemester 2016 abgeleiteten Codes werden in den zwei kognitiven Kategorien *Hypothese/Planung* (10 Subkategorien, 41 Codes) und *Auswertung* (6 Subkategorien, 26 Codes) sowie der sensomotorischen Kategorie *Durchführung* (8 Subkategorien, 44 Codes) zusammengefasst (Tab. 1).

Die Kodierung des audio-visuellen Datenmaterials erfolgt durch den Erstautor und zwei geschulte Projektmitarbeiter*innen. Die Interrater-Übereinstimmung der Kodierenden (20 % Doppeltkodierung) wird mit Hilfe des Übereinstimmungsmaßes Cohens Kappa (κ) innerhalb der drei Kategorien *Hypothese/Planung*, *Durchführung* und *Auswertung* überprüft. Die Übereinstimmungswerte der einzelnen Kategorien (Tab. 1) liegen im Bereich von 0.64 bis 0.96 und somit oberhalb des Grenzwertes für einen zufriedenstellenden Cohens Kappa ($\kappa > 0.60$) von WIRTZ & CASPAR (2002).

Tabelle 1: Übersicht der Kategorien und Subkategorien des Kodierleitfadens. Die Anzahl der jeweiligen Codes steht in Klammern.

Kategorien	**Hypothese / Planung**	**Durchführung**	**Auswertung**
Subkategorien (Anzahl der Codes)	Hypothese Formulierung (3) Nullhypothese (2) Weitere Hypothesen (2) Planung Stichprobe (3) Abhängige Variable (3) Unabhängige Variable (5) Störvariable (14) Messdesign (4) Wiederholungen Kontrollansatz (2)	Vorbereitung Proband (2) Einnahme Substanz (16) Störvariable (4) Experiment Proband (2) Kontrollansatz (2) Wiederholungen (2) Messdesign (2) Störvariable (14)	Darstellung der Daten (3) Beschreibung der Daten (2) Interpretation (2) Sicherheit (2) Methodenkritik (15) Ausblick (2)
Cohens κ (Interrater)	$\kappa = .64 - .91$	$\kappa = .81 - .96$	$\kappa = .76 - .94$

Nach Bildung eines Konsenses durch die Kodierenden werden die entstandenen Kodierungen in Performance-Scores umgewandelt (quantitative Auswertung qualitativer Daten: KUCKARTZ, 2014, S. 15). Die statistische Auswertung zur Untersuchung des Zusammenhangs zwischen den Ko-WADiS-Testwerten mit den Performance-Scores in den drei Kategorien des praktischen

Experimentiertests (F_1) erfolgt mit einer bivariaten Korrelationsanalyse im Programm SPSS (IBM, 2017). Dabei wird der Ko-WADiS-Test als eindimensionale Variable betrachtet (Krüger et al., 2020). Die berechneten SPEARMAN-Korrelationskoeffizienten ρ werden als Effektstärken nach COHEN (1988) interpretiert.

Zur Identifikation von Gruppenunterschieden (F_2) wird eine k-Means-Clusteranalyse (MACQUEEN, 1967) mit dem Ziel durchgeführt, die Proband*innen basierend auf ihren Performance-Scores derart in Gruppen (Cluster) zusammenzufassen, dass sich innerhalb eines Clusters Proband*innen mit möglichst homogenen Fähigkeiten und Fertigkeiten (Fälle) befinden, sich die einzelnen Cluster jedoch gegenseitig möglichst stark voneinander abgrenzen lassen (JANSSEN & LAATZ, 2017).

5 Ergebnisse

5.1 Vorhersagekraft des Ko-WADiS-Test (F_1)

Tabelle 2 zeigt die Maximalwerte, Mittelwerte, Standardabweichungen und Korrelationen der Performance Scores, aufgeteilt nach den drei Kategorien des Kodierleitfadens, miteinander sowie mit den erreichten Ko-WADiS-Testwerten.

Tabelle 2: Übersicht der deskriptiven Kennwerte und Interkorrelationen der kognitiven und sensomotorischen Performance-Scores mit den Ko-WADiS-Testwerten.

	N	MAX	*M*	SD	H/P	D	A	K
Hypothese/Planung (H/P)	30	37	15.16	4.85	–	.203	.260	.300
Durchführung (D)	30	13	7.52	1.70		–	-.331*	.104
Auswertung (A)	30	22	8.85	2.30			–	-.088
Ko-WADiS (K)	30	21	11.10	2.91				–

*. Die Korrelation ist auf dem Niveau von .05 (einseitig) signifikant.

Es zeigt sich, dass die Interkorrelationen der einzelnen Performance-Scores miteinander gering sind, abgesehen von der signifikanten, mittleren negativen Korrelation zwischen *Durchführung* und *Auswertung* (ρ = -.331, p = .037). Im Hinblick auf den Zusammenhang des praktischen Experimentiertests mit den Ko-WADiS-Testwerten lässt sich nur eine mittlere Korrelation mit der kognitiven

Kategorie *Hypothese/Planung* (ρ = .300, p = .054) erkennen. Die Kategorien *Durchführung* und *Auswertung* korrelieren nicht mit dem Paper-Pencil-Test.

5.2 Identifikation von Experimentiertypen (F_2)

Die k-Means-Clusteranalyse identifiziert vier voneinander abgrenzbare Fälle, von denen jeweils zwei Verläufe über die drei Kategorien des praktischen Experimentiertests annähernd parallel verlaufen und damit nur unterschiedliche Ausprägungen dokumentieren (Abb. 1a). Aus der Zusammenführung der parallelen Verläufe lassen sich zwei Experimentiertypen ableiten: Praktiker*innen und Allrounder. Die Praktiker*innen (n = 18) zeigen deutliche Stärken in der sensomotorischen Kategorie *Durchführung*; die Allrounder (n = 12) hingegen sind in allen drei Kategorien gleich stark, verglichen mit den Praktiker*innen aber stärker in den kognitiven Kategorien *Hypothese/Planung* und *Auswertung* (Abb. 1b).

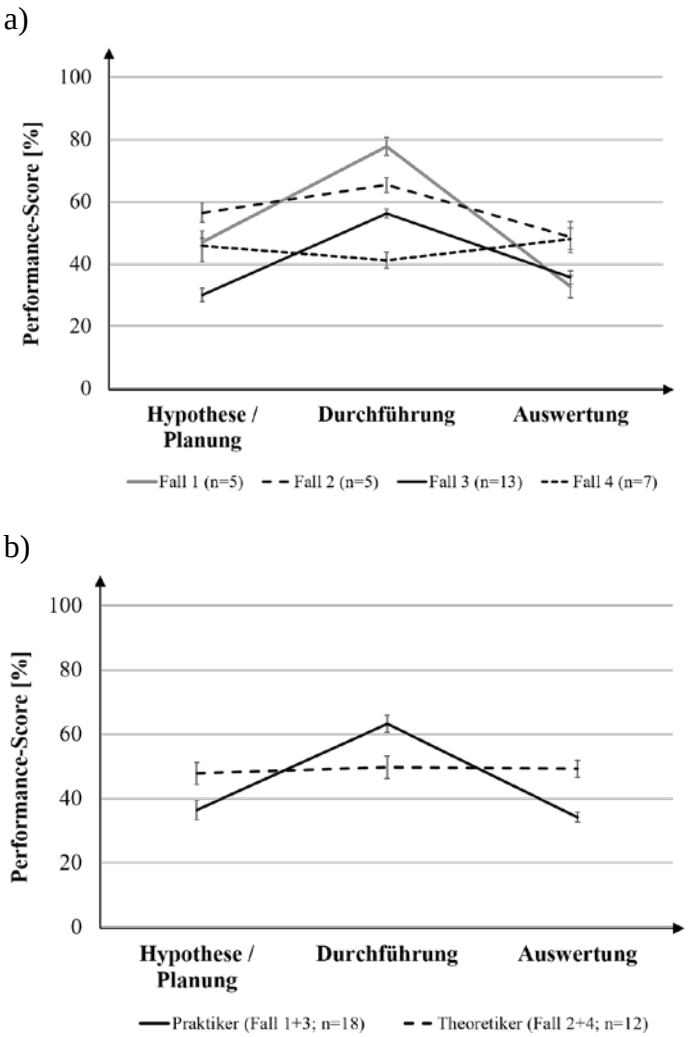

Abbildung 1: Prozentuale Darstellung der gemittelten Performance-Scores unter Berücksichtigung a) der identfizierten Fälle und b) der abgeleiteten Experimentiertypen.

Nach Aufteilung der Daten entsprechend der identifizierten Experimentiertypen wurde erneut die Interkorrelation der Performance-Scores mit den Ko-WADiS-Testwerten geprüft (Tab. 3).

Die Praktiker*innen zeigen innerhalb der Kategorien des praktischen Experimentiertests eine signifikante, starke Korrelation zwischen *Hypothese/ Planung* und *Durchführung* (ρ = .512, p = .030). Die kognitive Kategorie *Auswertung* korreliert nicht mit den anderen beiden Kategorien. Für die Kategorie *Hypothese/Planung* lässt sich eine mittlere Korrelation mit den Ko-WADiS-Testwerten erkennen (ρ = .441, p = .067), für die Kategorie *Auswertung* hingegen eine mittlere negative Korrelation (ρ = -.417, p = .085).

Für die Allrounder lässt sich eine mittlere Korrelation zwischen den Kategorien *Hypothese/Planung* und *Durchführung* identifizieren (ρ = .470, p = .123), sowie eine mittlere negative Korrelation zwischen *Durchführung* und *Auswertung* (ρ = -.479, p = .115). Ein Zusammenhang mit den Ko-WADiS-Testwerten besteht nicht.

Tabelle 3: Übersicht der Interkorrelationen nach Aufteilung der Proband*innen in die identifizierten Experimentiertypen: Praktiker*innen (n = 18) und Allrounder (n = 12).

	Praktikerînnen				**Allrounder**			
	H/P	D	A	K	H/P	D	A	K
Hypothese/Planung (H/P)	–	.512*	-.127	.441	–	.470	-.297	.154
Durchführung (D)		–	.005	.277		–	-.479	.076
Auswertung (A)			–	-.417			–	-.190
Ko-WADiS (K)				–				–

*. Die Korrelation ist auf dem Niveau von .05 (zweiseitig) signifikant.

6 Diskussion und Ausblick

Der Ausgangspunkt dieser Studie war die Frage, inwieweit die Performanz von Lehramtsstudierenden der Biologie durch die jeweils erreichten Ko-WADiS-Testwerte vorhergesagt werden können. Da Hammann et al. (2008) und Lawrence et al. (2001) keine oder nur geringe Zusammenhänge zwischen den Ergebnissen von schriftlichen MC-Kompetenztests und praktischen Experimentiertest für Schülerînnen zeigen konnten, wurde für die vorliegende Studie angenommen, dass dies auch für Studierende gilt und somit keine oder nur kleine Korrelationen zwischen beiden Datensätzen zu erwarten sind. Diese Annahme konnte bestätigt werden. Für die im Performance Assessment erhobenen kognitiven Fähigkeiten (*Hypothese/Planung, Auswertung*) und sensomotorischen Fertigkeiten

(*Durchführung*) bestehen nur kleine bis mittlere Korrelationen mit den KoWADiS-Testwerten. Es sollte jedoch berücksichtigt werden, dass die Performanz in praktischen Experimentiertests stark und individuell zwischen Aufgaben und Testzeitpunkten variiert (WEBB et al., 2000). Daher müsste eine relativ große Zahl an Experimentieraufgaben angeboten werden, um zu prüfen, ob neben persönlichen Faktoren auch Typ oder Kontext der Aufgabe die Performanz und damit den Zusammenhang mit schriftlichen Testergebnissen beeinflusst.

Durch die Anwendung einer k-Means-Clusteranalyse konnten im praktischen Experimentiertest vier Fälle identifiziert werden, die zusammengeführt zwei voneinander unterscheidbare Experimentiertypen mit unterschiedlichen Stärken in ihren kognitiven Fähigkeiten und sensomotorischen Fertigkeiten darstellen: Praktiker*innen und Allrounder. Dabei ist jedoch zu bedenken, dass das Ergebnis einer Clusteranalyse entscheidend von den verwendeten Analysedaten abhängt, d.h. die Auswahl der Variablen beeinflusst die Clusterbildung (JANSSEN & LAATZ, 2017). Um ausschließen zu können, dass es sich bei den identifizierten Experimentiertypen um einen möglichen Zufallsbefund handelt, wäre eine Fortführung der Studie mit einer größeren Stichprobe bzw. die Durchführung einer Clusteranalyse bei anderen praktischen Experimentiertests notwendig. Inwieweit die beiden Experimentiertypen in Anlehnung an KLAHR & DUNBAR (1988) auch andere Strategien zur Beantwortung einer Problemfrage nutzen, könnte Fragestellung einer Folgestudie sein.

Im nächsten Schritt werden die von den Proband*innen während des Experimentiertests angefertigten Versuchsdokumentationen hinsichtlich ihrer fachlichen Qualität untersucht. Damit soll untersucht werden, inwieweit sich der identifizierte schwache Zusammenhang zu den identifizierten Experimentiertypen weiter entwickelt (HILD et al., 2019).

Zitierte Literatur

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION [AERA, APA & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

ARNOLD, J., KREMER, K., & MAYER, J. (2013). Wissenschaftliches Denken beim Experimentieren – Kompetenzdiagnose in der Sekundarstufe II. In D. KRÜGER, A. UPMEIER ZU BELZEN, P. SCHMIEMANN, A. MÖLLER & D. ELSTER (Hrsg.), *Erkenntnisweg Biologiedidaktik 11* (S. 7-20). Kassel: Universitätsdruckerei.

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

BIERI BUSCHOR, C., & SCHULER BRAUNSCHWEIG, P. (2018). Predictive validity of a competence-based admission test − mentors' assessment of student teachers'occupational aptitude. *Assessment & Evaluation in Higher Education, 43*(4), 640-651.

BYBEE, R. W. (2002). Scientific Literacy – Mythos oder Realität. In W. GRÄBER, P. NENTWIG, T. KOBALLA & R. EVANS (Hrsg.), *Scientific Literacy. Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung* (S. 21-43). Opladen: Leske + Budrich.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

DR. DRESING & PEHL GMBH (2018). *f4transkript, Version 7*. Marburg: dr. dresing & pehl GmbH.

GIERE, R. N., BICKLE, J., & MAULDIN, R. F. (2006). *Understanding scientific reasoning*. Independence, KY: Wadsworth/Cengage Learning.

HAMMANN, M., PHAN, T. T. H., EHMER, M., & GRIMM, T. (2008). Assessing Pupils Skills in Experimentation. *Journal of Biological Education, 42*(2), 66-72.

HARTMANN, S., UPMEIER ZU BELZEN, A., KRÜGER, D., & PANT, H. (2015). Scientific reasoning in higher education. *Zeitschrift für Psychologie, 223*, 47–53.

HEIDRICH, J. (2017). *Erfassung von Experimentierkompetenz im universitären Kontext. Entwicklung und Validierung eines Experimentiertests zum Themenbereich Optik*. Dissertation.

HILD, P., GUT, C., & BRÜCKMANN, M. (2019). Validating performance assessments: measures that may help to evaluate students' expertise in 'doing science'. *Research in Science & Technological Education, 37*(4), 419-445.

IBM CORP. (2017). *IBM SPSS Statistics for Windows, Version 25.0*. Armonk, NY: IBM Corp.

JANSSEN, J., & LAATZ, W. (2017). Statistische Datenanalyse mit SPSS. Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests (9. Auflage). Berlin Springer.

KAMBACH, M. (2018). *Experimentierprozesse von Lehramtsstudierenden der Biologie – Eine Videostudie*. BIOLOGIE lernen und lehren, Bd. 23. Berlin Logos.

KLAHR, D., & DUNBAR, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1),1-48.

KRELL, M., REDMAN, C., MATHESIUS, S., KRÜGER, D., & VAN DRIEL, J. (2018). Assessing pre-service science teachers' scientific reasoning competencies. *Research in Science Education.*

KRÜGER, D., HARTMANN, S., NORDMEIER, V. & UPMEIER ZU BELZEN, A. (2020). Measuring Scientific Reasoning Competencies - Multiple Aspects of Validity. In O. ZLATKIN-TROITSCHANSKAIA, H. PANT, M. TOEPPER & C. LAUTENBACH (Hrsg.), *Student Learning in German Higher Education* (S. 261-280), Springer.

KUCKARTZ, U. (2014). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung*. (2. Auflage). Weinheim und Basel: Beltz Juventa.

LANDIS, J. R. & KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

LAWRENCE, F., HUFFMAN, D., & WELCH, W. (2001). The science achievement of various subgroups on alternative assessment formats. *Science Education, 85*(3), 279-290.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. LECAM & J. NEYMAN (Hrsg.), *Proceedings of the fifth Berkely symposium on mathematical statistics and probability* (Bd. 1, S. 281-297). Berkely, CA: University of California Press.

MAYER, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung* (S. 177-186). Berlin Springer.

MEIER, M., & MAYER, J. (2012). Experimentierkompetenz praktisch erfassen – Entwicklung und Validierung eines anwendungsbezogenen Aufgabendesigns. In U. HARMS & F. X. BOGNER (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik, Band 5* (S. 81-98). Innsbruck: StudienVerlag.

MÖLLER, A. & SPECHT, C. (2013). Reaktionsgeschwindigkeit beim Menschen. In P. SCHMIEMANN & J. MAYER (Hrsg.), *Experimentieren Sie! Biologieunterricht mit Aha-Effekt* (S. 60-62). Berlin Cornelsen.

ROTH, E. & HOLLING, H. (1999). *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (5. Auflage). München: R. Oldenbourg Verlag.

SCHECKER, H. & PARCHMANN, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 45-66.

SCHMIDT, D. (2016). *Modellierung experimenteller Kompetenzen sowie ihre Diagnostik und Förderung im Biologieunterricht*. BIOLOGIE lernen und lehren, Bd. 18. Berlin. Logos.

SCHMIEMANN, P., & LÜCKEN, M. (2014). Validität – Misst mein Test, was er soll? In D. KRÜGER, I. PARCHMANN & H. SCHECKER (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 107-118). Berlin, Heidelberg: Springer Spektrum.

SEKRETARIAT DER STÄNDIGEN KONFERENZ DER KULTUSMINISTER DER LÄNDER IN DER BUNDESREPUBLIK DEUTSCHLAND (KMK) (2019). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung* (Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 16.05.2019). Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (zuletzt abgerufen: 03.05.2020, 17.03 Uhr).

SEKRETARIAT DER STÄNDIGEN KONFERENZ DER KULTUSMINISTER DER LÄNDER IN DER BUNDESREPUBLIK DEUTSCHLAND (KMK) (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10)*. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf (zuletzt abgerufen: 03.05.2020, 17.05 Uhr).

SOLANO-FLORES, G. & SHAVELSON, R. J. (1997). Development of performance assessments in science: conceptual, practical and logistical issues. *Educational Measurement: Issues and Practice*, 16-25.

SURI, H. (2011). Purposeful sampling in qualitative research synthesis. *Qualitative research journal, 11*(2), 63-75.

TREADWAY, M. N. (2019). *Assessing student success: Predictive validity of the ACT science subscore* (Order No. 13812079, Eastern Michigan University). ProQuest Dissertations & Theses. Verfügbar unter https://search.proquest.com/docview/2229635354?accountid=11004 (zuletzt abgerufen: 03.05.2020, 17.12 Uhr)

UPMEIER ZU BELZEN, A. & KRÜGER, D. (2019). Modelle und Modellieren im Biologieunterricht: Ein Fall für Erkenntnisgewinnung. *Unterricht Chemie, 171*, 38-41.

VERBI SOFTWARE (2018). *MAXQDA – Software für qualitative Datenanalyse*. Berlin Consult. Sozialforschung GmbH.

WEBB, N. M., SCHLACKMAN, J., & SUGRUE, B. (2000). The dependability and interchangeability ofassessment methods in science. *Applied Measurement in Education,13*(3), 277-301.

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

WELLNITZ, N., & MAYER, J. (2016). Methoden der Erkenntnisgewinnung im Biologieunterricht. In A. SANDMANN & P. SCHMIEMANN (Hrsg.), *Erkenntnisse biologiedidaktischer Forschung. Schwerpunkte und Forschungsstände.* BIOLOGIE lernen und lehren, Bd. 1 (S. 61-82). Berlin: Logos.

WIRTZ, M., & CASPAR, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe & Huber.