Sabrina Mathesius/Moritz Krell/Annette Upmeier zu Belzen/Dirk Krüger

Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables[1]

Zusammenfassung: Im Projekt ValiDiS wird die Interpretation des Testwerts für den Multiple-Choice *Ko-WADiS*-Test zum *wissenschaftlichen Denken* unter Berücksichtigung des Validitätskriteriums *relations-to-other-variables* überprüft. Zur Untersuchung des empirischen Zusammenhangs des Testwerts mit den Konstrukten *schlussfolgerndes Denken* (verbale, numerische, figurale Intelligenz; *I-S-T 2000 R*) und *komplexes Problemlösen* (systematic exploration, system knowledge, control performance; *Genetics Lab*-Test) wurden etablierte Messinstrumente bei Biologie-Lehramtsstudierenden (N = 232) eingesetzt. Es zeigen sich positive, überwiegend signifikante Korrelationen mit mittleren Effekten, was die Annahme, dass es sich um distinkte Konstrukte mit verbindenden Kernfacetten handelt, stützt. Darüber hinaus besteht ein positiver Zusammenhang zwischen den Testwerten und der Abiturgesamtnote. Der Beitrag diskutiert die Befunde als Evidenz für eine valide Interpretation des *Ko-WADiS*-Testwerts.

Schlagworte: Testinstrument, Erkenntnisgewinnung, Problemlösen, Validität, Lehrkräfte

1. Einleitung

Innerhalb der Lehrkräftebildung ist das Professionswissen von Studierenden ein wichtiges Element der stetigen Evaluation sowie Optimierung der Lehramtsstudiengänge (vgl. Baumert & Kunter, 2006). Erst eine adäquate Ausbildung, welche das Professionswissen mit seinen Teilbereichen berücksichtigt, kann zukünftigen Lehrkräften ein erfolgreiches Unterrichten, Diagnostizieren und Fördern ihrer Lernenden ermöglichen. In dieser Studie wird als Professionswissen – angelehnt an die Unterteilung in pädagogisches, fachdidaktisches und fachliches Wissen – insbesondere die fachmethodische Ausbildung als Teil des Fachwissens betrachtet (vgl. Shulman, 1986). Im Zuge der reformierten Lehrkräftebildung in den Naturwissenschaften wurden zur Qualitätssicherung sowohl national als auch international zunehmend (Empfehlungen für) Aus-

1 Danksagung: Wir danken dem BMBF für die Förderung der Projekte Ko-WADiS/ValiDiS (01PK11004A/01PK15004A) innerhalb des Transferprojekts KoKoHs sowie der DFG für die Förderung des Projekts TypMoL (327507949). Unseren Proband*innen danken wir für ihre Teilnahme. Für ihre Unterstützung bei der Realisierung danken wir unseren Teammitgliedern: Tabea Dobberke, Anna Lena Engel, Maximilian Göhner, Maria-Elisa Puhlmann, Tom Röske und Contanze Rudwaleit.

bildungsstandards formuliert (vgl. z. B. KMK, 2019; VCAA, 2016). Demnach sollen Lehramtsstudierende fachspezifische Denk- und Arbeitsweisen wie beispielsweise das Experimentieren und Modellieren erlernen, diese in verschiedenen problemhaltigen Situationen anwenden und metaperspektivisch reflektieren können (vgl. Capps & Crawford, 2013).

Für die Erfassung des *wissenschaftlichen Denkens* bei Lehramtsstudierenden der naturwissenschaftlichen Fächer wurde der *Ko-WADiS*-Test entwickelt. Zur Untersuchung der Validität der Interpretation des Testwerts werden verschiedene Evidenzquellen herangezogen. In dieser Studie wird untersucht, welche empirischen Zusammenhänge zwischen dem *Ko-WADiS*-Testwert und den Konstrukten *komplexes Problemlösen* und *schlussfolgerndes Denken* sowie der Abiturgesamtnote als Hochschulzugangsberechtigung bestehen (Validitätskriterium: *relations-to-other-variables;* AERA, APA & NCME, 2014). Die Befunde werden als Hinweise einer validen Testwertinterpretation des *Ko-WADiS*-Tests diskutiert.

2. Theoretischer Rahmen

2.1 Validität

Validität ist neben Objektivität und Reliabilität ein Gütekriterium in der quantitativen Forschung. Dabei zielt die Analyse von Validität auf die mit Instrumenten generierten Testwerte und deren Interpretation. Kane (2013) fordert in seinem *argument-based approach to validation* die Konsequenzen einer Testung zu beachten und durch die Berücksichtigung verschiedener Evidenzquellen den argumentativen Nachweis zu erbringen, dass diese sowie die Testwertinterpretation legitim sind. Die *Standards for Educational and Psychological Testing* (AERA et al., 2014) beschreiben als Evidenzkriterien für die valide Testwertinterpretation neben dem operationalisierten *Testinhalt* die *interne Struktur* des Konstrukts und die *initiierten Antwortprozesse* während der Aufgabenbearbeitung sowie eine vergleichende Betrachtung des zu erfassenden Konstrukts mit externen Variablen *(relations-to-other-variables)*. Dieser Vergleich kann sich dabei sowohl auf die Betrachtung von unterschiedlichen Gruppen *(known-groups)* als auch auf unterschiedliche Konstrukte als externe Variablen beziehen. Letzteres berücksichtigt insbesondere, dass eine Konstruktdefinition auch durch das Aufzeigen von (fehlenden) Verbindungen zu anderen Konstrukten erfolgt. Durch die sich gegenseitig bedingende und aufeinander beziehende Betrachtung der Evidenzquellen zeigt sich die Untersuchung der Validität als ein Verfahren ohne Verwendung von Routinen. Inwiefern unter Berücksichtigung des Kriteriums *relations-to-other-variables* die Interpretation von Testwerten zum *wissenschaftlichen Denken* bestätigt oder falsifiziert werden kann, ist Gegenstand der vorliegenden Studie.

2.2 Wissenschaftliches Denken (WiDe)

Definition

Als Teilfacette der wissenschaftlichen Grundbildung ist *WiDe* (engl. *scientific reasoning*) definiert als Konstrukt, das sowohl die Fähigkeiten zum wissenschaftlichen Problemlösen als auch die Fähigkeiten zum metareflexiven Umgang mit dem Problemlöseprozess in den Naturwissenschaften beinhaltet (vgl. Lawson, 2004; Schwarz & White, 2005). Dabei sind kognitive Fähigkeiten wie deklaratives (Fachwissen sowie Wissen über naturwissenschaftliche Methoden), prozedurales (Wissen zur Ausführung naturwissenschaftlicher Methoden) und epistemologisches Wissen (Wissen bezogen auf die Natur der Naturwissenschaften) relevant (vgl. Gut-Glanzmann & Mayer, 2018; Osborne, 2013). Die naturwissenschaftlichen Problemlöseprozeduren gelten dabei für alle drei naturwissenschaftlichen Disziplinen. Sie sind in einer Vielzahl von Kompetenzmodellen abgebildet und betreffen naturwissenschaftliche Denkweisen wie das Entwickeln von Forschungsfragen, Aufstellen von Hypothesen, Planen von Untersuchungen und Auswerten von Daten (vgl. z. B. Klahr & Dunbar, 1988; Krell, 2018; Wellnitz et al., 2012) ebenso wie naturwissenschaftliche Arbeitsweisen (vgl. *styles of reasoning;* z. B. Beobachten, Experimentieren, Modellieren; Kind & Osborne, 2017). Diese Denk- und Arbeitsweisen besitzen spezifische Kompetenzanforderungen unter Berücksichtigung des deklarativen, prozeduralen und epistemologischen Wissens. Während das Methodenwissen disziplinübergreifend für die drei Naturwissenschaften gilt (vgl. Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010; Osborne, 2018), stellt sich das in Basiskonzepten strukturierte Fachwissen als disziplinspezifisch für die Naturwissenschaften Biologie, Chemie und Physik dar. Die Relevanz von fundiertem Fachwissen für naturwissenschaftliches Problemlösen wird international diskutiert (vgl. z. B. Osborne, 2018; Ruppert, Duncan & Chinn, 2017; Samarapungavan, 2018) und wird unter anderem damit erklärt, dass Fachwissen für die Konstruktion einer mentalen Repräsentation des Problems sowie für die Identifikation von zur Problemlösung relevanten Variablen notwendig ist (Fischer et al., 2014). Mayer, Sodian, Koerber & Schwippert (2014) konnten durch den Einsatz von papierbasierten Testinstrumenten und einer IRT-gestützten Auswertung in ihrer Studie für Grundschulkinder ($N = 155$) positive Zusammenhänge mit mittlerem Effekt zwischen Fähigkeiten der Planung und Auswertung von Untersuchungen (als Facetten des *scientific reasonings*) und allgemeinen kognitiven Fähigkeiten wie Intelligenz ($r = 0.38$, $p < .001$) und Lesekompetenzen ($r = 0.44$, $p < .001$) beschreiben. Eine Untersuchung von Zusammenhängen zwischen mentaler Kapazität und *scientific reasoning* bei Jugendlichen ($N = 210$) lässt ebenfalls positive Zusammenhänge mit mittlerem Effekt für die gemessenen Konstrukte erkennen ($r = 0.41$, $p < .01$; vgl. Kwon & Lawson, 2000). Dem folgend kann *WiDe* von allgemeinen kognitiven Fähigkeiten und logischen Denkprozessen abgegrenzt werden (vgl. Fischer et al., 2014; Mayer et al., 2014).

Erfassung mittels des Ko-WADiS-Tests

Mit Hilfe des *Ko-WADiS*-Tests sollen mit 123 Multiple-Choice-Aufgaben papierbasiert Fähigkeiten des *WiDe* für die Dimensionen *wissenschaftliches Untersuchen* und *wissenschaftliches Modellieren* bei Lehramtsstudierenden der Biologie, Chemie und Physik innerhalb einer Multi-Kohorten-Längsschnittstudie erfasst werden (Hartmann et al., 2015; Tab. 1). In den Aufgaben beurteilen die Proband*innen beispielsweise, inwiefern eine zu einem beschriebenen Phänomen vorgegebene Forschungsfrage empirisch überprüfbar, intersubjektiv nachvollziehbar, eindeutig sowie intern und extern konsistent ist oder es müssen Ursache-Wirkungs-Beziehungen im Umgang mit Variablen beim Experimentieren beurteilt werden (Mathesius, Upmeier zu Belzen & Krüger, 2014). Die problemhaltigen Situationen sind dabei für alle Aufgaben jeweils in einen biologischen, chemischen oder physikalischen Kontext eingebettet. Es liegt über die Testhefte hinweg eine Gleichverteilung für die zu untersuchenden Dimensionen sowie für die drei naturwissenschaftlichen Disziplinen vor. Für die Lösung der Aufgaben muss auf naturwissenschaftliches Methodenwissen zurückgegriffen werden, indem dieses in der jeweiligen problemhaltigen Situation zur Lösung angewendet wird (Mathesius, Upmeier zu Belzen & Krüger, 2018).

Die Validität der Testwertinterpretation wird für den *Ko-WADiS*-Test durch verschiedene Evidenzquellen evaluiert. Die qualitative Analyse von *Antwortprozessen* zeigt, dass das alleinige Lesen spezifischer Textstellen innerhalb der Aufgaben nicht zur Lösung ausreicht, sondern auf (prozedurales und epistemologisches) Methodenwissen zurückgegriffen werden muss (vgl. Mathesius et al., 2018). Untersuchungen zu spezifischen schwierigkeitserzeugenden Aufgabenmerkmalen zeigen, dass für die Aufgabenbearbeitung das Verständnis abstrakter Konzepte notwendig ist (vgl. Stiller et al., 2016). Des Weiteren lässt sich bezogen auf die *interne Struktur* von einem eindimensionalen Konstrukt *WiDe* sprechen (vgl. Hartmann et al., 2015; Mathesius et al., 2016). Die Reliabilität schwankt in Abhängigkeit der jeweils betrachteten Stichprobe, liegt aber insgesamt im zufriedenstellenden Bereich (z. B. $N = 2.247$, EAP/PV-Reliabilität = .54; vgl. Hartmann et al., 2015).

Dem Vergleich mit externen Variablen als Evidenzkriterium für die Betrachtung verschiedener Gruppen folgend wurden als Hinweise für eine valide Testwertinterpretation signifikant positive Zusammenhänge zwischen dem Testwert und der Studiendauer (Semesterzahl) von Lehramtsstudierenden gefunden. Zudem zeigen Lehramtsstudierende

Wissenschaftliches Denken		
Dimension	**Wissenschaftliches Untersuchen**	**Wissenschaftliches Modellieren**
Fähigkeiten	• Forschungsfragen formulieren (18) • Hypothesen generieren (16) • Untersuchungen planen (21) • Untersuchungen auswerten (17)	• Zweck von Modellen beurteilen (18) • Modelle testen (17) • Modelle ändern (13)

Tab. 1: Erfasste Fähigkeiten im Ko-WADiS-Test (vgl. Mathesius, Hartmann, Upmeier zu Belzen & Krüger, 2016). Anzahl der in dieser Studie eingesetzten Aufgaben in Klammern

mit zwei naturwissenschaftlichen Studienfächern einen höheren Testwert als Studierende mit nur einem naturwissenschaftlichen Studienfach (vgl. Hartmann et al., 2015; Mathesius et al., 2016).

Im Vergleich zu anderen Testinstrumenten in diesem Bereich, deren Validität als fragwürdig eingeschätzt wird (vgl. Opitz, Heene & Fischer, 2017; Osborne, 2013), liegen als besonderes Qualitätsmerkmal des *Ko-WADiS*-Tests bereits vielfältige Evidenzen für eine valide Testwertinterpretation vor. Es wurde bislang allerdings „noch nicht untersucht, ob der mutmaßliche Generalfaktor der Erkenntnisgewinnung nicht auf allgemeinere Fähigkeiten rückführbar ist. Hier wären beispielsweise allgemeine Intelligenz oder Problemlösefähigkeit denkbar“ (Hartmann et al., 2015, S. 53). Das Schließen dieser Forschungslücke wird in der vorliegenden Studie durch den vergleichenden Einsatz von etablierten Instrumenten angestrebt.

2.3 Schlussfolgerndes Denken (SlDe) als Facette von Intelligenz

Definition

Intelligenz wird als offenes Konstrukt verstanden, da eine allgemeingültige Definition fehlt. Konsens besteht darüber, dass kognitive Grundfähigkeiten das Ausführen von Aktivitäten erlauben, die unter anderem durch Komplexität, Abstraktion und Anpassungsfähigkeit beschreibbar sind. Dem folgend lassen sich kognitive Grundfähigkeiten als Fähigkeiten zum denkgestützten Lösen von dekontextualisierten Aufgaben und Problemen bezeichnen. *SlDe* umfasst dabei Fähigkeiten der verbalen, numerischen und figuralen Intelligenz als eine Facette der fluiden Intelligenz ohne wissensbezogene Intelligenzfähigkeiten (vgl. Liepmann, Beauducel, Brocke & Amthauer, 2007).

Erfassung mittels des I-S-T 2000 R

Der *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)* erfasst Teilfacetten von Intelligenz in verschiedenen Altersgruppen normiert. Er beinhaltet im Grundmodul jeweils 60 Aufgaben für verbale (Verb), numerische (Num) und figurale (Fig) Intelligenz, die in der Summe als Indikator für *SlDe* betrachtet werden (Liepmann et al., 2007).

Für den *I-S-T 2000 R* liegen vielfältige Evidenzen zur validen Testwertinterpretation vor. Explorative und konfirmatorische Faktorenanalysen bestätigen im Sinne der *Inhaltsvalidität* die zuvor operationalisierten Skalen des *SlDe* (Liepmann et al., 2007; dort als Konstruktvalidität). Der Vergleich mit anderen Intelligenz- sowie Aufmerksamkeitstests ergab Evidenz nach dem Kriterium *relations-to-other-variables* (Liepmann et al., 2007; dort als konvergente/diskriminante Validität). Der Einsatz in verschiedenen Altersgruppen liefert Hinweise für eine valide Testwertinterpretation im Sinne eines *known-groups*-Vergleichs (vgl. Steinmayer & Amelang, 2006; dort als Kriteriumsvalidität). Die festgestellte Reliabilität der Skalen ist hoch ($\alpha_{Verb} = .88$, $\alpha_{Num} = .95$, $\alpha_{Fig} = .87$; vgl. Liepmann et al., 2007).

2.4 Komplexes Problemlösen (KoPr)

Definition

Problemlösen kann definiert werden als „zielorientiertes Denken und Handeln in Situationen […], für deren Bewältigung keine routinierten Vorgehensweisen verfügbar sind" (Mayer, 2007, S. 178). Weitergehend spricht man „bei der Lösung von derartigen Problemen, die als dynamische Systeme hochgradig vernetzter Variablen beschrieben werden können, von *komplexem Problemlösen*" (Greiff & Fischer, 2013, S. 28). An den Lösungsprozess derartiger Aufgaben werden demnach hohe Anforderungen gestellt und die Erfassung des Konstrukts *KoPr* wird als Ergänzung zu traditionellen Intelligenztests diskutiert (vgl. Greiff & Fischer, 2013; Sonnleitner, Keller, Romain & Brunner, 2013).

Erfassung mittels des Genetics Lab-Tests

Der *Genetics Lab*-Test ist eine computerbasierte Mikrowelt zur Erfassung von *KoPr* bestehend aus zwölf in ihrer Komplexität zunehmenden Szenarien, in denen die Expression von Genen (unabhängige Variablen) fiktiver Kreaturen manipuliert werden kann und dies zu einem unterschiedlichen Phänotyp (abhängige Variable) führt (vgl. Sonnleitner et al., 2012). *KoPr* wird dabei in drei Aufgaben gegliedert: *Systematic exploration* (Expl): Es gilt, den Einfluss von Störvariablen zu identifizieren, unabhängige Variablen zu aktivieren und ihre ursächlichen Veränderungen zu identifizieren. Hierbei ist prozedurales Wissen über den systematischen Umgang mit Variablen relevant (z. B. nur eine Einflussvariable verändern). *System knowledge* (Knwl): Die Proband*innen dokumentieren als Abschluss der Exploration ein Modell des Systems. *Control performance* (Ctrl): Es muss von einem Ausgangszustand durch den planvollen Umgang mit den unabhängigen Variablen ein vorgegebener Zielzustand der abhängigen Variable erreicht werden.

Der Test wurde in verschiedenen Stichproben bei Schüler*innen der Klassenstufen neun bis elf eingesetzt (vgl. Sonnleitner et al., 2013). Für die Struktur des erfassten Konstrukts zeigten sich sowohl eine Aufteilung in die drei Dimensionen Expl, Knwl und Ctrl als auch die Interpretation als ein hierarchisches Konstrukt des komplexen Problemlösens mit spezifischen untergeordneten Komponenten als tragfähig. Die berichtete Reliabilität der drei Dimensionen ist hoch ($\alpha_{Expl} = .91$, $\alpha_{Knwl} = .90$, $\alpha_{Ctrl} = .79$; vgl. Sonnleitner et al., 2013; Evidenzkriterium *interne Struktur*). Im Sinne des Kriteriums *relations-to-other-variables* wird eine hohe bis mittlere Korrelation zwischen den Testwerten und der numerischen und figuralen Intelligenz (erfasst mit dem *I-S-T 2000 R;* vgl. Kapitel 2.5) als Evidenz für die valide Testwertinterpretation festgestellt (vgl. Sonnleitner et al., 2012, 2013).

2.5 Abiturgesamtnote als Hochschulzugangsberechtigung

Die Abiturgesamtnote gilt als vorrangiges Kriterium für die Auswahl von Studienbewerber*innen (vgl. Heine, Briedis, Didi, Haase & Trost, 2006), nicht zuletzt da ihr eine große Prädiktionsleistung auf den späteren Studienerfolg zugeschrieben wird und Schulnoten die Beurteilung von Fähigkeiten wie kognitive Leistungsfähigkeit, Leistungsmotivation und Lernbereitschaft widerspiegeln (vgl. Trapmann, Hell, Weigrand & Schuler, 2007). In der Abiturgesamtnote drücken sich schulfachliche Leistungen wie beispielsweise Fähigkeiten innerhalb des Kompetenzbereichs Erkenntnisgewinnung (vgl. KMK, 2005) in den drei naturwissenschaftlichen Disziplinen aus. Sie werden als Ergebnis von Prozessen des domänenspezifischen Wissenserwerbs und der Informationsverarbeitung aufgefasst. Ein positiver Zusammenhang von *KoPr* und Schulnoten wurde mehrfach festgestellt (vgl. z. B. Greiff & Fischer, 2013; Sonnleitner et al., 2013).

3. Forschungsanliegen

Im Sinne des *argument-based approach to validation* (vgl. Kane, 2013) ist in dieser Studie folgende Forschungsfrage leitend: Inwiefern liefert die vergleichende Betrachtung der Variablen *SlDe, KoPr* und Abiturgesamtnote im Sinne des Evidenzkriteriums *relations-to-other-variables* Hinweise für eine valide Testwertinterpretation des *Ko-WADiS*-Tests zur Erfassung von *WiDe* (vgl. AERA et al., 2014)?

Wie beschrieben, ist den drei Tests zu den Konstrukten *WiDe, SlDe* und *KoPr* gemein, dass die präsentierten Problemstellungen nicht allein durch einen Abruf von bekanntem Wissen lösbar sind. Die Proband*innen müssen jeweils Beziehungen und Sinnzusammenhänge zur Lösung der Testaufgaben herstellen und darüber hinaus (bewertende) Schlussfolgerungen ziehen. Dementsprechend werden verbindende Kernfacetten der betrachteten Konstrukte angenommen und es ergeben sich folgende Hypothesen:

Es besteht eine positive, signifikante Korrelation zwischen dem *Ko-WADiS*-Testwert und …

H1: … den drei Testwerten für die Skalen des *I-S-T 2000 R* mit mittlerer Effektstärke, da *WiDe* und *SlDe* allgemeine kognitive Fähigkeiten beinhalten und insbesondere verbale Intelligenz beim Lösen von textbasierten Aufgaben relevant ist (vgl. Liepmann et al., 2007; Stiller et al., 2016).

H2: … den drei Testwerten des *Genetics Lab*-Tests mit mindestens mittlerer Effektstärke, da *WiDe* und *KoPr* die Anwendung prozeduralen Wissens zum Umgang mit Variablen sowie Fähigkeiten im Modellieren erfordern (vgl. Mathesius et al., 2014; Sonnleitner et al., 2012).

H3: … der Abiturgesamtnote als Hochschulzugangsberechtigung mit mittlerer Effektstärke, da die Abiturgesamtnote als Indikator für ein breites Spektrum von Fähigkeiten gilt, wozu auch Fähigkeiten des *WiDe* im schulischen Kontext zählen (vgl. KMK, 2005).

H4: Aufgrund der in H1 bis H3 begründeten Zusammenhänge trägt die Berücksichtigung der Variablen *KoPr, SlDe* und Abiturgesamtnote insgesamt signifikant zur Aufklärung der Varianz des *Ko-WADiS*-Testwerts bei.

4. Methoden

4.1 Datenerhebung

Die beschriebenen Tests wurden an drei aufeinanderfolgenden Wochen unter Aufsicht eingesetzt. Für die Zusammenführung der Ergebnisse wurde von den Proband*innen ein anonymisierter Code generiert.

Der *Ko-WADiS*-Test mit 123 Aufgaben zu verschiedenen Kontexten der Disziplinen Biologie, Chemie und Physik wurde in Seminargruppen im Multimatrix-Design eingesetzt (20 verschiedene Testhefte mit im Mittel vergleichbarer Schwierigkeit; Bearbeitungszeit ca. 45 min; vgl. Hartmann et al., 2015). Die Testhefte sind dabei untereinander durch Blöcke von jeweils sieben Aufgaben verbunden. In jedem Testheft befinden sich Aufgaben aus jeder Disziplin für jede der sieben erfassten Fähigkeiten (vgl. Tab. 1). Aufgrund zu hoher MNSQ-Werte (> 1.3; vgl. Wright & Linacre, 1994) basiert die Datenauswertung hier auf 120 Aufgaben (Tab. 1). Die Personenfähigkeiten (WLE) wurden mit Hilfe des R Pakets TAM (vgl. Robitzsch, Kiefer & Wu, 2017) im einparametrisch-logistischen „Rasch"-Modell (1PLM) geschätzt.

Die Grundmodul-Kurzform des *I-S-T 2000 R* wurde als Gruppentest mit den Parallelformen A und B unter der vorgegebenen Testinstruktion durchgeführt (Bearbeitungszeit ca. 90 min; vgl. Liepmann et al., 2007). Als Personenfähigkeiten dienen Summenscores (Maximalwert von 60 für Teilfähigkeiten und 180 für *SlDe*).

Der *Genetics Lab*-Test wurde einzeln von den Proband*innen bearbeitet (Bearbeitungszeit ca. 60 min; Sonnleitner et al., 2012). Mit Hilfe des *Genetics-Lab-scoring-algorithm* (vgl. Keller & Sonnleitner, 2012) wurden Personenfähigkeiten für Expl (Wertebereich: 0–5), Knwl (Wertebereich: −0.46–0.54) und Ctrl (Wertebereich: 0–5) generiert.

Die Abiturgesamtnoten wurden auf einer Skala von 1 (= sehr gut) bis 4 (= ausreichend) über eine Selbstauskunft im Rahmen der Erfassung von demografischen Daten erhoben.

4.2 Stichprobe

An der Studie nahmen insgesamt $N = 232$ Biologie-Lehramtsstudierende im Rahmen einer Vollerhebung einzelner Studiensemester freiwillig teil (Tab. 2). Alle Proband*innen haben den *Ko-WADiS*-Test und jeweils eine Teilstichprobe hat den *I-S-T 2000 R* ($n = 169$) bzw. den *Genetics Lab*-Test ($n = 181$) bearbeitet. Die Teilstichproben ergeben sich dadurch, dass aus zeitökonomischen Gründen in drei Seminargruppen der *I-S-T-2000 R* und in zwei Seminargruppen der *Genetics-Lab*-Test nicht durchgeführt werden konnte.

4.3 Datenauswertung

Zur Hypothesenprüfung wurde eine bivariate Korrelationsanalyse (Pearson) mit den Werten der drei Tests sowie der Abiturgesamtnote (H1–H3) durchgeführt. In Bezug auf H4 wurde eine lineare Regressionsanalyse mit dem *Ko-WADiS*-Testwert als abhängiger Variable und den in H1–H3 betrachteten Variablen als unabhängigen Variablen durchgeführt.

5. Ergebnisse

5.1 Kennwerte der Tests

Ko-WADiS: Die MNSQ-Werte der 120 Items deuten einen ausreichend guten Item-Fit an ($.72 < \text{MNSQ}_{\text{Infit}} < 1.22$; $.59 < \text{MNSQ}_{\text{Outfit}} < 1.29$). Die Reliabilität liegt im erwarteten Bereich und die Lösungswahrscheinlichkeit sowie Aufgabenschwierigkeit zeigen eine angemessene Testschwierigkeit (Tab. 3). Ergebnisse vorheriger Studien, welche einen positiven, signifikanten Zusammenhang zwischen dem Testwert und der Anzahl der naturwissenschaftlichen Studienfächer sowie der Anzahl der studierten Hochschulsemester zeigen (vgl. Hartmann et al., 2015; Mathesius et al., 2016), konnten repliziert werden ($M \pm SD_{\text{WLE\_zwei nawi Fächer}} = .36 \pm .54$, $M \pm SD_{\text{WLE\_ein nawi Fach}} = .12 \pm .76$; $p = .018$, $d = .42$; kleiner Effekt; bivariate Pearson-Korrelation $r_{\text{Hochschulsemester}} = .24$, $p < .001$).

I-S-T 2000 R: Die durchschnittlichen Personenfähigkeiten (Tab. 4) entsprechen den berichteten Normwerten 21- bis 25-jähriger Gymnasiast*innen (vgl. Liepmann et al., 2007; *SlDe:* $M \pm SD = 113.90 \pm 21.31$; Verb: $M \pm SD = 39.80 \pm 7.52$; Num: $M \pm SD = 42.29 \pm 10.69$; $M \pm SD = 34.40 \pm 8.19$). Die Reliabilitäten für die Skalen zur numerischen und figuralen Intelligenz fallen deutlich höher als für die Skala zur verbalen Intelligenz aus (Tab. 4).

Genetics Lab: Die Reliabilitäten fallen für alle drei Skalen hoch aus, die Personenfähigkeiten liegen im erwarteten Bereich (Tab. 5; vgl. Sonnleitner et al., 2013).

Studiengang	Bachelor: *n* = 98	Master: *n* = 134
Hochschulsemester	*M* = 8.17, *SD* = 4.39	
Geschlecht	weiblich: *n* = 161	männlich: *n* = 69
Alter	*M* = 24.45, *SD* = 4.45	17–49 Jahre
Anzahl naturwissenschaftlicher Studienfächer	eins: *n* = 192	zwei: *n* = 40

Tab. 2: Informationen zur Stichprobe

	N	**Anzahl Items**	**Lösungswahrscheinlichkeit**		**Aufgabenschwierigkeit (Item-Parameter im 1PLM)**		**Reliabilität**
			M ± SD	**Range**	**M ± SD**	**Range**	**(EAP)**
Ko-WADiS	232	120	.48 ± .16	.05–.84	.03 ± .88	−2.19–1.99	.55

Tab. 3: Kennwerte des Ko-WADiS-Tests

	n	**Anzahl Items**	**Personenfähigkeit**		**Reliabilität**
			M ± SD	**Range**	**(α)**
SIDe	168	180	112.99 ± 18.71	48–158	.91
Verb	168	60	39.64 ± 5.64	9–49	.64
Num	168	60	39.24 ± 10.74	7–59	.92
Fig	168	60	33.95 ± 7.52	15–50	.81

Tab. 4: Kennwerte des I-S-T 2000 R

	N	**Anzahl Items**	**Personenfähigkeiten**		**Reliabilität**
			M ± SD	**Range**	**(α)**
Expl	181	12	3.37 ± 1.41	.00–5.00	.93
Knwl	181	12	.39 ± .14	−.03–.536	.92
Ctrl	181	12	3.74 ± .92	.44–5.00	.84

Tab. 5: Kennwerte des Genetics Lab-Tests

Abiturgesamtnote: Die durchschnittliche Abiturgesamtnote der hier betrachteten Stichprobe (N = 226; $M \pm SD$ = 2.09 ± .53) entspricht den Daten der *Ko-WADiS*-Gesamtstichprobe innerhalb der Multi-Kohorten-Längsschnittstudie (N = 3415, davon haben n = 443 die Abiturgesamtnote mehrfach berichtet; $M \pm SD$ = 2.10 ± .53) sowie den bundesweiten Angaben zum Durchschnitt der Abiturgesamtnote (vgl. Ramm, Multrus & Bargel, 2011).

5.2 Hypothesenprüfung: Korrelationsanalyse (H1–H3)

Der *Ko-WADiS*-Testwert hängt am stärksten mit dem Wert der Verb-Skala des *I-S-T-2000 R* zusammen (Tab. 6). Der Zusammenhang zwischen dem *Ko-WADiS*-Testwert und dem Wert der Fig-Skala ist hierbei signifikant kleiner als der zwischen dem *Ko-WADiS*-Testwert und den Werten der Verb-Skala (z = 2.76, p = .003, q = .27; kleiner Effekt) sowie der Num-Skala (z = 2.61, p = .005, q = .21; kleiner Effekt). Für die drei Skalen des *Genetics Lab*-Tests ergeben sich untereinander keine signifikanten Korrelationsunterschiede zum *Ko-WADiS*-Testwert. Der Zusammenhang zwischen dem *Ko-WADiS*-Testwert und der Abiturgesamtnote ist mit kleinem Effekt signifikant positiv.

5.3 Hypothesenprüfung: Regressionsanalyse (H4)

Der Wert der Durban-Watson-Statistik liegt bei d = 2.21 und zeigt damit an, dass nicht von einer zu großen Autokorrelation der Residuen ausgegangen werden muss. Die Kollinearitätsdiagnose ergibt Toleranzwerte ≥ .12 und spricht damit gegen eine Multikollinearität der unabhängigen Variablen (Urban & Mayerl, 2011). In der Regressionsanalyse werden etwa 24 % der Varianz der abhängigen Variablen durch die berücksichtigten unabhängigen Variablen erklärt (Tab. 7).

		I-S-T 2000 R				*Genetics Lab*			Abiturgesamtnote
		SIDe	Verb	Num	Fig	Expl	Knwl	Ctrl	
Ko-WADiS	*r*	.44	.44	.39	.20	.37	.33	.40	.26
	p	< .001	< .001	< .001	.009	< .001	< .001	< .001	.019
	n	169	169	169	169	181	181	181	225

Tab. 6: Bivariate Korrelationen (Pearson) zwischen dem Ko-WADiS-Testwert (WLE), den I-S-T 2000 R-Testwerten (Summenscore) und den Genetics Lab-Testwerten (Personenfähigkeiten) sowie der Abiturgesamtnote

	b	*se(b)*	β	*p*
(Konstante)	−1.68	0.55	–	.003
Verb	0.03	0.01	0.26	.005
Num	0.01	0.06	0.19	.061
Fig	0.00	0.08	0.02	.987
Expl	0.12	0.11	0.24	.307
Knwl	−1.35	1.16	−0.27	.245
Ctrl	0.19	0.09	0.25	.036
Abiturgesamtnote	−0.07	0.11	−0.05	.532

Anmerkung: Angegeben sind das Regressionsgewicht *b,* dessen Standardfehler *se*(*b*), der standardisierte Regressionskoeffizient β sowie der Signifikanzwert *p*. Methode: Einschluss. Korrigiertes R^2 = .24.

Tab. 7: Lineare Regressionsanalyse mit dem Ko-WADiS-Testwert (WLE) als abhängiger Variable

6. Diskussion

6.1 Generelle Limitierungen der Studie

Bevor die Ergebnisse bezogen auf das Forschungsanliegen diskutiert werden, gilt es, einige methodische und inhaltliche Einschränkungen zu berücksichtigen.

Es handelt sich bei der vorliegenden Studie um eine Untersuchung auf Populationsebene, weshalb keine Aussagen auf Ebene einzelner Proband*innen erfolgen können. Demnach ist die geringe Reliabilität für den *Ko-WADiS*-Test insbesondere mit Verweis auf eine IRT-gestützte Auswertung nicht in dem Umfang als Maß für die Güte des Messinstruments zu verstehen, wie dies für Messinstrumente, die im Rahmen der klassischen Testtheorie konzipiert und ausgewertet werden, der Fall ist (vgl. Adams, 2005). Ein Einfluss auf die Lösung der Aufgaben sowohl durch die präsentierten Kontexte als auch durch die Reihenfolge der Aufgaben in den Testheften kann nicht ausgeschlossen werden. Bei der Zusammenstellung der Testhefte wurde daher im Vorfeld auf eine im Mittel vergleichbare Schwierigkeit geachtet (vgl. 4.1 Datenerhebung).

Die untersuchten Gelegenheitsstichproben entstammen verschiedenen Seminargruppen. Es erfolgte grundsätzliche eine Vollerhebung innerhalb der Seminare. Dennoch kann dabei aufgrund der Freiwilligkeit der Teilnahme sowie der anonymen Datenerhebung nicht ausgeschlossen werden, dass eine positive Selektion vorliegt (vgl. Nagy, 2005; Schachtschneider, 2016). Die große Spannweite der Lösungswahrscheinlichkeit sowie der Personenfähigkeit lassen auf eine ausreichend heterogene Stichprobe schließen (vgl. Tab. 3, 4, 5). Darüber hinaus muss einschränkend berücksichtigt werden, dass alle Proband*innen denselben Studiengang an derselben Universität absolvieren. Für eine umfassendere Einordnung der Ergebnisse wäre die Stichprobe um weitere univer-

sitäre Standorte sowie um Studierende eines naturwissenschaftlichen Faches ohne Lehramtsoption zu ergänzen (vgl. Freyer, Epple, Brand, Schiebner & Sumfleth, 2014).

Um einen erhöhten *cognitive load* bei den Proband*innen zu vermeiden, wurden die drei Testinstrumente nicht an einem Termin eingesetzt. Infolgedessen ist nicht auszuschließen, dass Einflussfaktoren wie Interesse, Motivation und Selbstkonzept nicht nur individuell, sondern auch je Erhebungstermin als Störvariablen die Beantwortung der Testaufgaben unterschiedlich beeinflusst haben. Für die Kontrolle solcher Einflüsse wird vorgeschlagen, Messinstrumente parallel einzusetzen (vgl. Nehring, 2014; Ryan & Deci, 2000). Darauf wurde hier verzichtet, weil die Zeitkontingente für entsprechende Testungen fehlten.

Unterschiede in den Testformaten und den Operationalisierungen der zu messenden Konstrukte können als unterscheidende Faktoren wirken, wobei sich insbesondere die Erhebung der *Genetics Lab*-Testwerte durch die Dynamik und die Erstellung von neuem Wissen von den anderen papierbasierten Formaten abgrenzt (Sonnleitner et al., 2012).

6.2 Kennwerte der Tests

Die Testwerte der drei Instrumente (*Ko-WADiS*-Test, *I-S-T 2000 R, Genetics Lab*-Test) zeigen im Vergleich zu publizierten Normwerten keine auffälligen Abweichungen (vgl. Hartmann et al., 2015; Liepmann et al, 2007; Mathesius et al., 2016; Sonnleitner et al., 2013). Die Reliabilitäten liegen in den erwarteten Bereichen. Darüber hinaus liegen die MNSQ-Werte für 120 Aufgaben des *Ko-WADiS*-Tests im zufriedenstellenden Bereich. Es konnte bestätigt werden, dass Studierende eines hohen Fachsemesters ebenso wie Studierende mit zwei naturwissenschaftlichen Studienfächern einen höheren Testwert im *Ko-WADiS*-Test erzielen als Studierende eines niedrigen Fachsemesters bzw. nur eines naturwissenschaftlichen Studienfachs (vgl. Hartmann et al., 2015; Mathesius et al., 2016). Da die *I-S-T 2000 R*-Testwerte Übereinstimmungen mit der Referenz-Normstichprobe zeigen, ist auch hier davon auszugehen, dass diese als *SlDe* mit den Teilfacetten Verb, Num und Fig interpretiert werden dürfen (vgl. Liepmann et al., 2007). Trotz veränderter Zielstichprobe (Studierende anstatt Schüler*innen) zeigen sich für die *Genetics Lab*-Testwerte keine Deckeneffekte, so dass insgesamt davon auszugehen ist, dass diese für das Konstrukt *KoPr* valide interpretiert werden können (vgl. Sonnleitner et al., 2013).

6.3 Hypothesenprüfung

Es zeigen sich zwischen dem *Ko-WADiS*-Testwert und den *I-S-T 2000 R*-Testwerten sowie den *Genetics Lab*-Testwerten überwiegend signifikante, positive Korrelationen mit mittleren Effekten (Tab. 6). Dies stützt die Annahme, dass es sich um distinkte Konstrukte mit verbindenden Kernfacetten handelt und steht somit im Einklang mit anderen Studien im Bereich des *WiDe* (vgl. z. B. Kwon & Lawson, 2000; Mayer et al.,

2014). Die Zusammenhänge zwischen dem *Ko-WADiS*-Testwert und den *Genetics Lab*-Testwerten deuten auf die postulierten Gemeinsamkeiten in der Anwendung von prozeduralem Wissen zum Umgang mit Variablen – beispielsweise bezogen auf Ursache-Wirkungs-Beziehungen beim Experimentieren – hin (vgl. Mayer, 2007). Spezifische Varianz kann durch die Operationalisierung der Aufgaben im *Ko-WADiS*-Test sowie durch die unterschiedlichen Testformate erklärt werden. Die Aufgaben zur Erhebung von *WiDe* folgen dabei dem Ansatz, naturwissenschaftliches Methodenwissen diziplinübergreifend zu erfassen.

Der *Ko-WADiS*-Testwert hängt am stärksten mit den Werten der Verb-Skala des *I-S-T 2000 R* zusammen. Gründe hierfür können in der verbalen Informationspräsentation innerhalb der Multiple Choice-Aufgaben des *Ko-WADiS*-Tests liegen (Mathesius et al., 2014). Vorherige Studien liefern Hinweise, dass zur Aufgabenbearbeitung im *Ko-WADiS*-Test neben einem generellen Leseverständnis insbesondere Fähigkeiten des *WiDe* notwendig sind (Mathesius et al., 2018; Stiller et al., 2016). Zur Unterstützung dieser Annahme hätte ein Testinstrument zur Erfassung des generellen Leseverständnisses in das Versuchssetting integriert werden müssen (vgl. Mayer et al., 2014). Der geringste Zusammenhang wird zwischen dem *Ko-WADiS*-Testwert und dem Wert für die Fig-Skala festgestellt. Dies lässt sich durch die Operationalisierung der Aufgaben des *Ko-WADiS*-Tests erklären, welche nur im geringen Maße auszuwertende Grafiken enthalten (vgl. Stiller et al., 2016).

Der signifikante Zusammenhang zwischen dem *Ko-WADiS*-Testwert sowie der Abiturgesamtnote stützt die Position, dass schulische Leistungen Differenzen innerhalb der Testwerte in Leistungstests erklären können (vgl. Klusmann, Trautwein, Lüdtke, Kunter & Baumert, 2009). Das Auswahlverfahren der Studierenden über den Numerus Clausus bedingt dabei eine selektierte Stichprobe – hier basierend auf der Zulassung zu einem Biologie-Lehramtsstudium. Die Aussagekraft von selbst berichteten Schulnoten wird in einigen Studien kritisch diskutiert, wenngleich eine hohe Korrelation zwischen berichteten und tatsächlichen Noten gefunden wurde (vgl. Trapmann et al., 2007). Für die hier befragte Stichprobe, welche eine Teilstichprobe der Multi-Kohorten-Längsschnitt-Studie innerhalb des Gesamtprojekts darstellt, kann bei 95.04 % der mehrfach erhobenen Werte eine konsistente Angabe der Abiturgesamtnote festgestellt werden, so dass den Angaben für die vergleichenden Analysen Vertrauen geschenkt wird.

Der Zusammenhang zwischen dem *Ko-WADiS*-Testwert sowie den verschiedenen Einflussvariablen kann durch die Regressionsanalyse beschrieben werden (Tab. 7). Durch die gemeinsame Betrachtung der unabhängigen Variablen zeigen nur noch die Werte für Verb und Ctrl eine signifikante Prädiktionsleistung für den *Ko-WADiS*-Testwert. Die Werte der Ctrl-Skala spiegeln dabei in besonderem Maße die Verschränkung von hypothetisch-deduktivem Arbeiten während des Experimentierens und Modellierens wider, worin der besondere Zusammenhang zum *Ko-WADiS*-Testwert gesehen werden kann (Hartmann et al., 2015; vgl. Sonnleitner et al., 2013). Dies stützt die Annahme der operationalisierten distinkten Konstrukte mit verbindenden Kernfacetten.

Zusammenfassend liefern die Ergebnisse für das Validitätskriterium *relations-to-other-variables* Hinweise, die Interpretation des *Ko-WADiS*-Testwerts im Sinne des

Konstrukts *WiDe* zu anderen distinkten Konstrukten wie dem *KoPr* oder *SlDe* abgrenzend vorzunehmen. Die beiden letzteren Konstrukte unterscheiden sich vom ersten dabei insbesondere dahingehend, dass sie kein naturwissenschaftliches Methodenwissen erfordern, wodurch sich die Divergenz der drei Konstrukte erklären lässt. Entsprechend der theoriegeleiteten und systematischen Konstruktionsanleitung wird in den Aufgaben des *Ko-WADiS*-Tests das zur Bearbeitung benötigte Fachwissen im Aufgabenstamm präsentiert und allein das naturwissenschaftliche Methodenwissen soll zur Lösung in den Fokus rücken. Studien zur Untersuchung des Lösungsprozesses mittels der Methode des Lauten Denkens sowie Eyetracking-Daten stützen diese Annahme (vgl. Mathesius et al., 2018).

In der Regressionsanalyse können 76 % der Varianz nicht durch die innerhalb dieser Studie berücksichtigten Variablen aufgeklärt werden. Die verbleibende Varianz könnte als spezifische Wissensbasis zur naturwissenschaftlichen Problemlösung (d. h. deklaratives, prozedurales und epistemologisches Methodenwissen; vgl. Gut-Glanzmann & Mayer, 2018; Osborne, 2013) und damit als spezifisch für den *Ko-WADiS*-Testwert gesehen werden.

7. Ausblick

Weitere Studien sollten sich zur Aufklärung ebenso wie zur Untersuchung der kausalen Zusammenhänge zwischen den betrachteten Konstrukten anschließen.

Des Weiteren wäre die Varianzaufklärung durch die Erhebung von deklarativem Fachwissen zu den Aufgabenkontexten des *Ko-WADiS*-Tests zu erweitern (vgl. Mayer, 2007). Ergänzend zur Note der Hochschulzugangsberechtigung könnte zudem als Kriterium das Maß des Studienerfolgs beispielsweise durch die Erhebung von Studiennoten für ausgewählte Module modelliert werden (vgl. Schachtschneider, 2016). Diese Werte können zum *Ko-WADiS*-Testwert (im Studienverlauf) in Beziehung gesetzt werden, um hierdurch den Zusammenhang der verschiedenen Konstrukte auch längsschnittlich modellieren zu können (vgl. Trapmann et al., 2007). Wiederholte Messzeitpunkte sollten dafür in den Blick genommen werden, um die Stabilität des Konstrukts *SlDe* im Vergleich zu einer potentiellen Entwicklung für die Konstrukte *WiDe* und *KoPr* zu prüfen. Hierfür liegen für das Konstrukt *WiDe* Projektdaten der längsschnittlichen Erhebung vor (vgl. Mathesius et al., 2016). Ergänzende Hinweise zur Varianzaufklärung werden durch den durchgeführten Vergleich mit einem weiteren Testinstrument zum *WiDe* erwartet (vgl. Großschedl, Mahler, Kleickmann, & Harms, 2014).

Der gefundene Zusammenhang zwischen dem *Ko-WADiS*-Testwert und den *I-S-T* 2000 *R*-Testwerten wird zukünftig auch für die Auswahl von Proband*innen zur vergleichenden Untersuchung ihrer theoretischen und praktischen Fähigkeiten im Bereich des *WiDe* genutzt (vgl. Göhner & Krell, 2018).

Literatur

Adams, R. J. (2005). Reliability as a Measurement Design Effect. *Studies in Educational Evaluation, 31*(2-3), 162–172.

AERA, APA & NCME = American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

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

Capps, D., & Crawford, B. (2013). Inquiry-Based Professional Development: What does it take to support teachers in learning about inquiry and nature of science? *International Journal of Science Education, 35*(12), 1947–1978.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific Reasoning and Argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45.

Freyer, K., Epple, M., Brand, M., Schiebener, J., & Sumfleth, E. (2014). Studienerfolgsprognose bei Erstsemesterstudierenden in Chemie. *Zeitschrift für Didaktik der Naturwissenschaften, 20*(1), 129–142.

Greiff, S., & Fischer, A. (2013). Der Nutzen einer komplexen Problemlösekompetenz. *Zeitschrift für Pädagogische Psychologie, 27*(1-2), 27–39.

Großschedl, J., Mahler, D., Kleickmann, T., & Harms, U. (2014). Content-Related Knowledge of Biology Teachers from Secondary schools: Structure and learning opportunities. *International Journal of Science Education, 36*(14), 2335–2366.

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

Göhner, M., & Krell, M. (2018). Modellierungsprozesse von Lehramtsstudierenden der Biologie. *Erkenntnisweg Biologiedidaktik, 17,* 45–61.

Hartmann, S., Mathesius, S., Stiller, J., Straube, P., Krüger, D., & Upmeier zu Belzen, A. (2015). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte. Das Projekt Ko-WADiS. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* (S. 39–58). Bad Heilbrunn: Klinkhardt.

Heine, C., Briedis, K., Didi, H.-J., Haase, C., & Trost, G. (2006). *Auswahl- und Eignungsfeststellungsverfahren.* Hannover: HIS Hochschul-Informations-System GmbH.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73.

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

Keller, U., & Sonnleitner, P. (2012). Genetics Lab Scoring Algorithm. http://hdl.handle.net/10993/13098 [28. 10. 2018].

Kind, P., & Osborne, J. (2017). Styles of Scientific Reasoning. *Science Education, 101*(1), 8–31.

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

Klusmann, U., Trautwein, U., Lüdtke, O., Kunter, M., & Baumert, J. (2009). Eingangsvoraussetzungen beim Studienbeginn. *Zeitschrift für Pädagogische Psychologie, 23*(3-4), 265–278.

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD (Hrsg.) (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss.* München & Neuwied: Wolters Kluwer.

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD (Hrsg.) (2019). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung.* Berlin. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf [22. 05. 2019].

Krell, M. (2018). Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie. *Zeitschrift für Didaktik der Naturwissenschaften, 24*(1), 1–15.

Kwon, Y., & Lawson, A. E. (2000). Linking Brain Growth with the Development of Scientific Reasoning Ability and Conceptual Change during Adolescence. *Journal of Research in Science Teaching, 37*(1), 44–62.

Lawson, A. (2004). The Nature and Development of Scientific Reasoning. *International Journal of Science and Mathematics Education, 2*(3), 307–338.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R).* Göttingen: Hogrefe.

Mathesius, S., Hartmann, S., Upmeier zu Belzen, A., & Krüger, D. (2016). Scientific Reasoning as an Aspect of Pre-service Biology Teacher Education. Assessing competencies using a paper-pencil test. In T. Tal & A. Yarden (Hrsg.), *The Future of Biology Education Research. A selection of papers presented at the 10th conference of European Researchers in Didactics of Biology (ERIDOB)* (S. 93–110). Haifa, Israel.

Mathesius, S., Upmeier zu Belzen, A., & Krüger, D. (2014). Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. Entwicklung eines Testinstruments. *Erkenntnisweg Biologiedidaktik, 13,* 73–88.

Mathesius, S., Upmeier zu Belzen, A., & Krüger, D. (2018). Eyetracking als Methode zur Untersuchung von Multiple-Choice-Aufgaben zum wissenschaftlichen Denken. In M. Hammann & M. Lindner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (Bd. 8, S. 225–244). Innsbruck/Wien/Bozen: Studienverlag.

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

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific Reasoning in Elementary School Children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29,* 43–55.

Nagy, G. (2005). *Berufliche Interessen, kognitive und fachgebundene Kompetenzen.* Dissertation. Freie Universität Berlin. Verfügbar unter http://www.diss.fu-berlin.de/cgi-bin/zip.cgi/2007/109/Fub-diss2007109.zip [07. 05. 2019].

Nehring, A. (2014). *Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung* (Studien zum Physik- und Chemielernen, Bd. 177). Dissertation. Berlin: Logos Verlag.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring Scientific Reasoning: A review of test instruments. *Educational Research and Evaluation, 23*(3-4), 78–101.

Osborne, J. (2013). The 21st Century Challenge for Science Education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10,* 265–279.

Osborne, J. (2018). Styles of Scientific Reasoning: What can we learn from looking at the product, not the process, of scientific reasoning? In F. Fischer, C. Chinn, K. Engelmann & J. Osborne (Hrsg.), *Scientific Reasoning and Argumentation. The roles of domain-specific and domain-general knowledge* (S. 162–186). New York/London: Routledge.

Ramm, M., Multrus, F., & Bargel, T. (2011). *Studentensituation und studentische Orientierungen.* Bonn: BMBF.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). Package ‚TAM'. https://cran.opencpu.org/web/packages/TAM/TAM.pdf [07.05.2019].

Ruppert, J., Duncan, R., & Chinn, C. (2017). Disentangling the Role of Domain-Specific Knowledge in Student Modeling. *Research in Science Education, 49*(3), 921–948.

Ryan, R.M., & Deci, E.L. (2000). Self-Detetermination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist, 55*(1), 68–78.

Samarapungavan, A. (2018). Construing Scientific Evidence: The role of disciplinary knowledge in reasoning with and about evidence in scientific practice. In F. Fischer, C. Chinn, K. Engelmann & J. Osborne (Hrsg.), *Scientific Reasoning and Argumentation. The roles of domain-specific and domain-general knowledge* (S. 66–86). New York/London: Routledge.

Schwarz, C., & White, B. (2005). Metamodeling Knowledge. *Cognition and Instruction, 23*(2), 165–205.

Shulman, L. (1986). Those Who Understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Schachtschneider, Y. (2016). *Studieneingangsvoraussetzungen und Studienerfolg im Fach Biologie* (Biologie lernen und lehren, Bd. 12). Dissertation. Berlin: Logos Verlag.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab. *Psychological Test and Assessment Modeling, 54*(1), 54–72.

Sonnleitner, P., Keller, U., Romain, M., & Brunner, M. (2013). Students' Complex Problem-solving Abilities. *Intelligence, 41*(5), 289–305.

Steinmayr, R., & Amelang, M. (2006). Erste Untersuchungen zur Kriteriums-Validität des I-S-T 2000 R an Erwachsenen beiderlei Geschlechts. *Diagnostica, 52*(4), 181–188.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing Scientific Reasoning. A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education, 41*(5), 721–732.

Trapmann, S., Hell, B., Weigrand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs. *Zeitschrift für Pädagogische Psychologie, 21*(1), 11–27.

Urban, D., & Mayerl, J. (2011). *Regressionsanalyse.* Wiesbaden: VS.

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.

Wright, B., & Linacre, J. (1994). Reasonable Mean-square Fit Values. *Rasch Measurement Transactions, 8*(3), 370.

VCAA = Victorian Curriculum and Assessment Authority (2016). *Victorian Certificate of Education. Biology.* Melbourne, Victoria: VCAA.

Abstract: For the ValiDiS project, the interpretation of the test score of a multiple-choice test for scientific reasoning (the Ko-WADiS-test) is examined, taking into account the validity criterion relations-to-other-variables. In order to investigate the empirical relation of the test score with deductive reasoning (verbal, numerical, figurative intelligence, I-S-T 2000 R) and complex problem solving (systematic exploration, system knowledge, control performance, Genetics Lab-test), established assessment instruments were used to obtain test scores of pre-service science teachers (N = 232). There are positive, mostly significant correlations with medium effects, which supports the assumption that they are distinct constructs with linking facets. In addition, there is a positive correlation between the test score and the final grade. In this contribution, the findings are discussed as evidence for a valid test score interpretation of the Ko-WADiS-test.

Keywords: Assessment Tool, Scientific Inquiry, Problem Solving, Validity, Teachers

Anschrift der Autor*innen

Sabrina Mathesius, Freie Universität Berlin,
Didaktik der Biologie,
14195 Berlin, Deutschland
E-Mail: sabrina.mathesius@fu-berlin.de

Dr. Moritz Krell, Freie Universität Berlin,
Didaktik der Biologie,
14195 Berlin, Deutschland
E-Mail: moritz.krell@fu-berlin.de

Prof. Dr. Annette Upmeier zu Belzen, Humboldt-Universität zu Berlin,
Fachdidaktik und Lehr-/Lernforschung Biologie,
10099 Berlin, Deutschland
E-Mail: annette.upmeier@biologie.hu-berlin.de

Prof. Dr. Dirk Krüger, Freie Universität Berlin,
Didaktik der Biologie,
14195 Berlin, Deutschland
E-Mail: dirk.krueger@fu-berlin.de