Konzeption und Evaluation von universell designten Lernumgebungen und Assessments zur Förderung und Erfassung von Nature of Science Konzepten

Von der Naturwissenschaftlichen Fakultät
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation
von
Malte Walkowiak, M.Ed.

2019

Referent: Prof. Dr. Andreas Nehring
Koreferentin: Prof. Dr. Insa Melle
Tag der Promotion: 26.03.2019

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kurzfassung

Die vorliegende Arbeit thematisiert die Herausforderungen, die sich durch sozialpolitische (Migration) und bildungspolitische Einflüsse (Inklusion) in den letzten Jahren für den naturwissenschaftlichen Unterricht in Deutschland ergeben haben. Durch Entscheidungen dieser Art in der jüngeren Vergangenheit steigt die Vielfalt der Lernendenmerkmale in Klassen an. Für sich genommen ist diese Entwicklung nicht problematisch. Da jedoch in Deutschland das Abschneiden in Schulleistungstests von diversen Lernendenmerkmalen bestimmt ist, führen neue Dimensionen, wie der Migrationsstatus oder aber ein sonderpädagogischer Unterstützungsbedarf zu weiteren Differenzlinien. Zusätzlich führen neue Schulformen dazu, dass es Bewegungen innerhalb der SchülerInnenschaft an diese neuen Schulformen oder an das Gymnasium gibt.

Universal Design for Learning (UDL) ist ein möglicher Ansatz, um diesen Anforderungen gerecht zu werden. UDL vermeidet traditionelle One-size-fits-all-Ansätze für das Lehren und Lernen. Hierfür bietet UDL eine ganze Reihe von Ideen für unterrichtliche Adaptionen, die darauf abzielen, die Partizipation des Lernenden am Unterricht zu erhöhen. Mit Blick auf die Diversität der SchülerInnenschaft versucht UDL, den Gedanken der Barrierefreiheit im Lehren und Lernen zu etablieren.

Das Universal Design for Assessment (UDA) verfolgt in gleicherweise den Gedanken, dass traditionelle Assessments unzureichend sind, um der Diversität der ProbandInnen gerecht zu werden. Durch Testadaptionen versucht UDA, die konstruktirrelevante Varianz (KIV) zu reduzieren und dadurch die Testzugänglichkeit zu steigern.

Nature of Science (NOS) stellt eine wesentliche Komponente innerhalb der Naturwissenschaftsdidaktik dar und ist Teil der naturwissenschaftlichen Grundbildung. Außerdem ist NOS Bestandteil von vielen curricularen Vorgaben. Aufgrund der herausragenden Bedeutung von NOS und aufgrund der Tatsache, dass es weder im Feld von UDA noch von UDL bisher umgesetzt wurde, stellt es den Inhaltsbereich dieser Arbeit dar.

In einer qualitativen Studie wurden die UDL-Lernumgebungen konzipiert und evaluiert. Eine zweite quantitative Studie vergleicht das konzipierte UDA-basierte Assessment mit dem publizierten, unveränderten Assessment sowie zwei Lernumgebungen, die sich im Grad der UDL-Implementierung unterscheiden.

Sowohl in der qualitativen als auch in der quantitativen Studie waren die UDL-Lernumgebungen in der Lage, NOS-Konzepte auch in kurzer Zeit zu elaborieren. Der Grad der UDL-Implementierung ist dabei weniger entscheidend. Außerdem unterstützen die Daten aus beiden Assessmentformen der quantitativen Studie zumindest partial-skalare Messinvarianz. Somit sind die Werte aus beiden Assessments auf der Mittelwertsebene vergleichbar. Allerdings sind die psychometrischen Kennwerte des UDA-Assessments im Vergleich zum Originalassessment meist besser. Zusammenfassend stellen sowohl UDL als auch UDA sowie der verwendete Untersuchungsgang geeignete Mittel dar, um die Effekte von Interventionen im Feld von NOS und Diversität zu untersuchen.

Schlagworte: Diversität - Universal Design - Messinvarianz

Summary

The present study addresses the challenges that have arisen due to social-political (migration) and educational-political (inclusive teaching) influences in recent years for science education in Germany. Those decisions increased the variety of learner characteristics in classes. By itself, this development is not problematic. However, as the performance of school achievement tests in Germany is determined by various learner characteristics, new dimensions, such as the status of migration or for a special-needs education, lead to further differences. In addition, new types of schools are leading to movements within the student body to these new school forms or to the Gymnasium.

Universal Design for Learning (UDL) is a possible approach to meet these requirements. UDL avoids traditional one-size-fits-all approaches to teaching and learning. For this purpose, UDL offers a whole range of ideas for teaching adaptations, which aim to increase the learner's participation in class. With regard to students's diversity, UDL tries to establish the concept of accessibility in teaching and learning.

Similarly, the Universal Design for Assessment (UDA) pursues the idea that traditional assessments are inadequate to accommodate the diversity of the test-takers. Through test accomodations, UDA attempts to reduce the construct-irrelevant variance (CIV) and thereby increase test accessibility.

Nature of Science (NOS) is an essential component of science education and is part of basic science education. In addition, NOS is part of many curricular requirements. Due to the outstanding importance of NOS and the fact that it has not yet been implemented in the field of UDA or UDL, it is the content of this work.

In a qualitative study, the UDL learning environments were designed and evaluated. A second quantitative study compares the designed UDA-based assessment with the published, unchanged assessment and two learning environments, which differ in the degree of UDL implementation.

In both the qualitative and the quantitative study, the UDL learning environments were able to elaborate NOS concepts in a short time. The degree of UDL implementation is less critical. In addition, the data from both assessment forms of the quantitative study support at least partial scalar measurement invariance. Thus, both assessments are comparable at the mean level. However, the psychometric parameters of the UDA assessment are usually better in comparison to the original assessment. In summary, both UDL and UDA, as well as the methods the study uses, are appropriate tools to study NOS field interventions and diversity.

Key words: Diversity - Universal Design - Measurement invariance

1 | Einleitung

Bisher gibt es in der Chemiedidaktik wenige konzeptionelle Antworten auf die Herausforderungen, die durch die Einführung eines inklusiven Unterrichtes entstanden sind. Darüber hinaus sind Lernendenleistungen in Deutschland weiterhin u.a. von der Lesefähigkeit, dem sozioökonomischen Status oder dem Migrationshintergrund abhängig (Autorengruppe Bildungsberichterstattung, 2016, 2018). Vor dem Hintergrund eines breiten Inklusionsverständnisses, was die *Diversität von Lernenden* nicht ein- (ausschließlich ein sonderpädagogischer Unterstützungsbedarf), sondern multidimensional (zusätzlich z.B. die Lesefähigkeit) versteht, ergibt sich daher die Frage nach gänzlich neuen unterrichtlichen Konzeptionen.

Mit dem Universal Design for Learning (UDL) (Abschnitt 2.3.1) besteht ein theoretisches Rahmenwerk zur Konzeption von Unterricht, welches die *Zugänglichkeit* zu Lerninhalten thematisiert. *Zugänglichkeit* wird hier im Sinne von *Barrierefreiheit* gedacht. Die Grundannahme von UDL besteht darin, dass jeder monomodale Unterricht (z.B. bezüglich des Lerninhalts oder der Verarbeitung dessen) Barrieren enthält. Multimodalität einer Lernumgebung wird durch multiple Repräsentationsformen, multiple Verarbeitungsformen und motivierende bzw. motivationserhaltende Elemente in der Lernumgebung hergestellt (CAST, 2018). Bildlich gesprochen stellt UDL *das Was*, *das Wie* und *das Warum des Lernens* in den Vordergrund der Unterrichtsplanung. Durch die breite Konzeption von UDL kann es zur Gestaltung von Unterricht in allen Unterrichtsfächern verwendet werden. Bisherige Metastudien zu UDL zeigen, dass vor allem das *Prinzip der multiplen Repräsentationsformen* umgesetzt wurde (Al-Azawei, Serenelli & Lundqvist, 2016). Außerdem wird vorwiegend mit qualitativen Methoden gearbeitet. Damit ergibt sich eine Lücke bezüglich quantitativer Interventionen, die die drei UDL-Prinzipien und nicht nur einzelne implementieren.

Außerdem berichten Studien oftmals von einer positiven Wahrnehmung des eigenen Lernprozesses durch die Probanden sowie der Unterrichtsmaterialien (Rao, Ok & Bryant, 2014). Es wird hingegen kaum über die Lernzuwächse in Studien berichtet (Capp, 2017). Es kann hiervon ausgehend vermutet werden, dass Assessments oft nicht adäquat genug sind, um Effekte, verursacht durch die Lernumgebungen, abzubilden. Unter Umständen sind die Barrieren von Assessments selbst zu groß. In der Folge steigt die *konstruktirrelevante Varianz* (KIV) und überlagert somit mögliche statistische Effekte. Mit dem Universal Design for Assessment (UDA) (Abschnitt 2.3.2) besteht ein theoretischer Rahmen, Assessments vor dem Gedanken der Barrierefreiheit zu gestalten (Thompson, Johnstone & Thurlow, 2002). UDA diskutiert hierzu die *Testzugänglichkeit*, die sich darin ausdrückt, dass die Probandin/der Proband ausschließlich mit dem Zielkonstrukt und den dazu notwendigen internen Charakteristika interagiert (Beddow, 2011). In dieser Weise dient UDA der Reduzierung der KIV.

Als Inhaltsbereich für die Studie wurde *Nature of Science* (NOS) ausgewählt. NOS stellt eine wesentliche Komponente von *naturwissenschaftlicher Grundbildung* dar und ist Teil von vielen curricularen Standards (Bernholt, Neumann & Nentwig, 2012; NGSS Lead States, 2013). NOS umfasst unter anderem die datenbasierte Argumentation, die Kritik am aktuellen Wissen und betont damit dessen Vorläufigkeitscharakter. NOS umfasst aber auch die Frage nach den Beteiligten der Wissensproduktion (Abschnitt 2.4). Diese Aufzählung ist notwendigerweise nicht vollständig. In der Summe können die Dimensionen von NOS als reflexive Praxis verstanden werden, die die charakteristischen Grundzüge naturwissenschaftlicher Erkenntnisgewinnung und die Eigenschaften naturwissenschaftlichen Wissens diskutiert (Höttecke, 2001; Kremer, 2010).

Bisher wurde NOS nicht in UDL-Interventionen umgesetzt. Typischerweise wird NOS darüber hinaus in Interviewsituationen mit qualitativen Instrumenten erhoben. Um die Lücke von quantitativen Studien im Feld von UDL zu adressieren, wurde ein kontextloses Likertinstrument zur Messung von NOS-Konzepten verwendet (Urhahne, Kremer & Mayer, 2008). Um das Problem der KIV zu adressieren, wurde das Assessment in seiner letzten publizierten Form und in einer nach dem UDA adaptierten Version verwendet (Kampa, Neumann, Heitmann & Kremer, 2016). Für die Studie wurden UDL-basierte Lernumgebungen (Abschnitte 7.1 und 7.2) zur Förderung von Konzepten zur *Rechtfertigung von naturwissenschaftlichem Wissen* entwickelt und qualitativ wie quantitativ evaluiert.

Alle Analysen wurden in R (Version 3.5.2) vorgenommen (R Core Team, 2016). Die Dissertation selbst wurde mit `markdown` (Allaire et al., 2016) kombiniert mit `bookdown` (Xie, 2016), `knitr` (Xie, 2018) und `kableExtra` (Zhu, 2018) verfasst. Im Anhang finden sich zwei komplette `lavaan`-Skripte (Rosseel, 2012). Eines zur longitudinalen Messinvarianzprüfung sowie ein anderes zu einem latenten Wachstumsmodell. Beide sind im Multi-Group-Ansatz spezifiziert. Außerdem wird die Bestimmung der internen Konsistenz via `semTools` und `psych` gezeigt (Jorgensen et al., 2018; Revelle, 2018). Zuletzt findet sich im Anhang noch ein Skript zur DIF-Analyse mit `pairwise` (Heine, 2017).

2 | Theoretischer Rahmen

2.1 Umgang mit Lernendenmerkmalen im Unterricht

Im Diskurs zum Umgang mit Merkmalen von Schülerinnen und Schülern im Unterricht wird oftmals von der *Heterogenität* der Lernenden gesprochen (Menthe & Sander, 2016; Oser & Blömeke, 2012; Ve Wittig, 2014). Wenn die Diskussion auf mögliche sonderpädagogische Unterstützungsbedarfe fokussiert, wird oft der Begriff *Inklusion* verwendet (Werning & Baumert, 2013). Um einem ganzheitlichen Blick auf Lernendenmerkmale zu erhalten, bietet es sich jedoch an, von *Diversität* zu sprechen (Sliwka, 2012). Jedoch stellt bereits die Wahl des Begriffs ein Indiz dafür dar, aus welcher Perspektive der Umgang mit Lernendenmerkmalen gepflegt wird.

Durch gesellschaftliche und politische Prozesse besteht eine aktuelle Herausforderung für Lehrerinnen und Lehrer aber auch für die Fachdidaktiken darin, der wachsenden Unterschiedlichkeit von Lernenden zu begegnen (Schimank, 2015). Dabei gibt es zwei Perspektiven: eine negative und eine positive. Die *negative Perspektive* sieht in der wachsenden Unterschiedlichkeit von Schülerinnen und Schülern (z.B. hinsichtlich ihrer persönlichen Eigenschaften, ihres sozialen oder ökonomischen Hintergrundes) eine Erschwernis für den Unterricht (Altrichter, Trautmann, Wischer, Sommerauer & Doppler, 2009; Markic & Abels, 2014). Mit dieser Sichtweise ist der Begriff *Heterogenität* verknüpft, weil die *Abweichung von einer Norm* betont wird. Die *positive Sichtweise* sieht vor allem im Begriff *Diversität* den Abbau von Diskriminierungen, die aufgrund verschiedener persönlicher Merkmale (z.B.: Geschlecht oder Hautfarbe) entstehen können (Krell, 2013). Diese Sichtweise lässt sich mit der Forderung *Lernen und Lehren für Alle umreißen*. Beide Perspektiven werden im Folgenden diskutiert.

2.1.1 Negative Perspektive: Die Heterogenität von Lernenden

Heterogenität ist ein Begriff mit Bezug auf Lerngruppen. Der Ausdruck bezeichnet eine Differenz bezüglich eines gemeinsamen Merkmals (z.B. nach Alter, Geschlecht, Intelligenz) (Ve Wittig, 2014). Die Ausprägung dieses Merkmals variiert über alle Gruppenmitglieder. Diesem sehr praktischen Zugang folgen jedoch zwei Probleme. Oft ist völlig unklar, welches Merkmal gemeint ist, wenn von *heterogenen Klassen* gesprochen wird. Diese Beschreibung erfolgt auf sozialkonstruierten Differenzlinien basierend auf persönlichen Erfahrungen. Diese Erfahrung bestimmt weiterhin die Art und Weise der Merkmalsausprägung. Ein Beispiel

hierfür ist das Geschlecht, was im europäischen Kulturraum dichotom, andernorts jedoch polytom gedacht und verstanden wird.

> Schließlich "kann es durchaus sein, dass eine Klasse im Hinblick auf ein bestimmtes Merkmal als homogen eingestuft wird, sich allerdings in weiteren Merkmalen als heterogen beschreiben lässt" (Ve Wittig, 2014, S. 14).

Es wird jedoch deutlich, dass der Begriff *Heterogenität* mit der Merkmalsauswahl und der sich daraus ergebenden Zuschreibung variiert. Zusammenfassend stellt *Heterogenität* immer eine normative und unter Umständen willkürliche Beschreibung dar. Damit stellt *Heterogenität* eine räumlich und zeitlich begrenzte Zustandsbeschreibung dar. Etwas als heterogen zu beschreiben, kann sich daher auch mit der Betrachtungsweise und der eigenen Erfahrung ändern (Trautmann & Wischer, 2011).

Außerdem ist die Frage nach der/den relevanten Heterogenitätsdimension/en für die Gestaltung von Lehr-Lern-Prozessen bisher jedoch weitestgehend unbeantwortet (Trautmann & Wischer, 2011). Vermutlich lässt sich eine solche Frage aber auch nicht allgemeingültig beantworten. Trautmann & Wischer (2011) schlagen dennoch zwei konzeptionelle Vorgehensweisen für die Beantwortung dieser Frage vor. *Lehr-Lernpsychologische Zugänge* wollen Lernendenmerkmale identifizieren, die für die Gestaltung von Unterricht relevant sind. So können auch relevante Dimensionen über Modelle zu *Gutem Unterricht* identifiziert werden, die gewissermaßen als Voraussetzung betrachtet werden können. Hierzu zählen auch die Merkmale der Lehrerinnen und Lehrer. Im Anschluss an das Angebots-Nutzungsmodell von Helmke (2009) leiten Trautmann & Wischer (2011) vier im Zusammenhang mit Schulleistungen stehende Heterogenitätsdimensionen ab. Erstens weisen alle Schülerinnen und Schüler *Vorwissen* im Hinblick auf den jeweiligen Lerngegenstand auf. Als zweites werden *kognitive Grundfähigkeiten bzw. Intelligenz* angeführt. Allerdings sind diese problematisch, weil *kognitive Grundfähigkeiten* einerseits als Konstrukt einschlägig, aber auch umstritten ist, weil das Konstrukt als "statisch, eindimensional und zu deterministisch gedacht" wird (Trautmann & Wischer, 2011, S. 45). Drittens werden *motivationale und affektive Merkmale* angeführt. Diese wirken auf das Lernverhalten und damit indirekt auf die Leistung von Schülerinnen und Schülern, wenn man Lernen als einen durch den Lernenden selbstgesteuerten Prozess begreift. Gleichwohl gibt es eine ganze Reihe von Konstrukten (z.B. Motivation, Freude, schulisches oder fachliches Selbstkonzept) von denen eine solche Wirkung bekannt ist bzw. angenommen werden kann. Entsprechend liegen vermutlich komplexe Wechselwirkungen zwischen diversen Konstrukten vor. Viertens stellen die *strukturellen Bedingungsfaktoren* von Lernenden eine Voraussetzung für Schulleistungen dar. Das prominenteste Beispiel ist der sozioökonomische Status. Dieser steht immer wieder im engen Zusammenhang mit Schulerfolg (Autorengruppe Bildungsberichterstattung, 2016, 2018; Helmke & Weinert, 1997). Gerade hier erweist sich die *negative Perspektive* auf Merkmale von Schülerinnen und Schülern als problematisch. Es kann schnell zu kurzschlüssigen Erklärungen kommen.

> "Dabei muss [...] darauf hingewiesen werden, dass *Sozialschicht* eine bildungssoziologische Kategorie ist, die für sich genommen *keinen direkten Erklärungswert* hat. Das Leistungsniveau eines Kindes [...] ist nicht deshalb niedriger, weil es zur sozial niedrigeren Schicht gehört, sondern weil der kognitive Anregungsgehalt, die elterlichen Standards und Erwartungen, ihre leistungsbezogenen Erklärungen und Sanktionen und ihr eigenes Engagement für Schulleistungen des Kindes in niedrigeren sozialen Schichten typischerweise geringer

> ausgeprägt sind. Dies sind die hinter der *Sozialschicht* liegenden eigentlichen Wirkfaktoren" (Helmke, 2012, S. 57, Herv. im Orig.).

Die von Helmke (2012) angesprochenen Wirkfaktoren bilden ihrerseits aber vermutlich komplexe Geflechte und Wechselwirkungen. Eine einseitige Betrachtung von Heterogenität kann damit zu unmittelbaren Benachteiligungspraxen führen.

Der zweite von Trautmann & Wischer (2011) vorgeschlagene Zugang leitet sich aus einer Gesellschafts- und Erkenntnistheorie ab.

> "Anders als bei der Lehr-Lern-Forschung, die sich für Lerner- bzw. Gruppenmerkmale letztlich nur in Bezug auf die Optimierung von Unterrichtsprozessen (und Steigerung von Lernleistungen) interessiert, liegt der Ausgangspunkt hier bei gesellschaftlichen (Ungleichheits-)Verhältnissen und deren Herstellung und Fortschreibung, auch im und durch das Bildungswesen [entgegenzuwirken]" (Trautmann & Wischer, 2011, S. 47).

Hieraus ergeben sich eine ganze Reihe von Differenzlinien (Geschlechter, Lernbehinderung oder ethnische Zugehörigkeit usw.). Während jedoch jede dieser Heterogenitätsdimensionen für sich eine spezielle Ausrichtung des Unterrichts notwendig erscheinen lässt, ist dies mit Bezug auf Unterrichten zu kritisieren. Und zwar allein deshalb, weil alle Menschen an vielen Differenzlinien gleichzeitig positioniert sind. Eine reduktionistische Sichtweise ist daher unangebracht, weil sie der Komplexität von Unterricht nicht gerecht wird.

Zusammenfassend muss daher aus der *negativen Perspektive*, wenn von zunehmenden heterogenen Klassen gesprochen wird, gefragt werden: Heterogen - mit Bezug auf was und warum? Denn ohne konkrete Antworten auf diese Frage, ist eine reflektierte Unterrichtsgestaltung nicht möglich.

2.1.2 Positive Perspektive: Die Diversität von Lernenden

Durch änderungen auf politischer und gesellschaftlicher Ebene ist es in den letzten Jahrzehnten immer wieder in Deutschland zu Veränderungen im Bildungssystem gekommen. Hierzu zählen unter anderem neue Schulformen im Sekundarbereich, die durch Auflösung und Zusammenfassung von Haupt- und Realschulen entstanden sind. Diese stellen neben den Gymnasien nun eine neue Säule im deutschen Schulsystem dar. Als ein Resultat steigt auch die Vielfalt der Schülerschaft in den Gymnasien, wenn Schülerinnen und Schüler aus dem oberen Leistungsbereich mit hoher Bildungsaspiration aus diesen Schulformen hinzukommen. Außerdem führt die Ratifizierung der UN-Behindertenrechtskonvention 2009 (UN-BRK) zu einem gemeinsamen Lernen von Lernenden mit und ohne sonderpädagogischen Unterstützungsbedarf. Schließlich steigt auch durch Zuwanderung der Anteil von Schülerinnen und Schülern mit Migrationshintergrund in allen deutschen Bildungseinrichtungen (Fischer, Rott & Veber, 2014). Der Begriff *Diversität* nimmt aber, im Gegensatz zum Begriff *Heterogenität*, eine andere Position zur Unterschiedlichkeit von Schülerinnen und Schülern ein (Fischer et al., 2014; Lee, Miller & Januszyk, 2015). So werden die Merkmale von Schülerinnen und Schülern nicht als bloße Voraussetzung für Lehr-Lern-Prozesse, sondern vielmehr als Ressource für diese verstanden. Die positive Perspektive sieht in der Unterschiedlichkeit der Schülerinnen und Schüler nicht mehr ein Problem, sondern Normalität.

> "Die Diversität der Individuen hinsichtlich ihrer herkunftsbedingten Sozialisation, ihren ethnischen und religiösen Wurzeln, ihrer Begabungsprofile und Interessen innerhalb einer Schule kann dann zu einer Lernressource werden, wenn dazu im Unterricht und in der Organisation einer Schule die notwendigen Voraussetzungen geschaffen werden" (Sliwka, 2012, S. 171).

Sliwka (2012) sieht in diesem Anspruch aber keine plötzliche änderung, sondern das Ergebnis eines Prozesses, den Schule und der in Schulen stattfindende Unterricht durchlaufen müssen (Abb. 2.1).

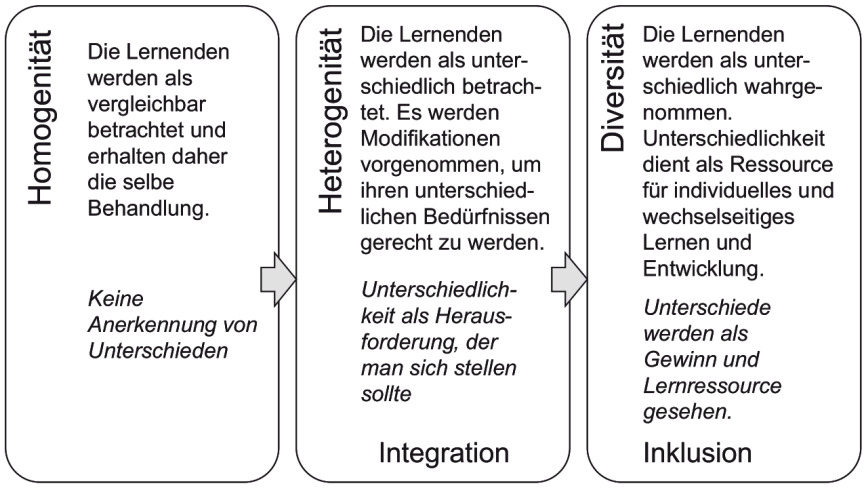


Abbildung 2.1: Von der Homogenität zur Diversität (Sliwka, 2012, 171).

In diesem Zusammenhang diskutiert Sliwka (2012) Begriffspaare: Während *Heterogenität* maximal einen Zustand der *Integration* darstellen kann, weil lediglich ein Umgang mit Lernendenmerkmalen stattfindet, stellt *Inklusion* den Zustand der Anerkennung und der *Wertschätzung* dar (Brownell, Smith, Crockett & Griffin, 2012; Sliwka, 2012). Damit wird deutlich, dass ein Unterricht mit dieser Anlage, Lehrerinnen und Lehrer mit bestimmten Haltungen und Einstellungen bedarf (Tab. 2.1). Es ist daher nicht verwunderlich, dass ein *diversitätswertschätzender Unterricht* mit *Einstellungen und Haltungen von Lehrenden* zusammenhängt (Kahn & Lewis, 2014; Savolainen, Engelbrecht, Nel & Malinen, 2012; Sharma & Sokal, 2015; Specht et al., 2016).

Es wird an dieser Stelle deutlich, was unter der Aussage *Lernendenmerkmale als Ressource begreifen* zu verstehen ist: Setzt ein Unterricht die genannten Kriterien um, tritt die Schülerin und der Schüler in den Mittelpunkt des Unterrichtsgeschehens, indem seine/ihre Weltsicht aufgenommen wird und die Unterrichtsplanung nach diesen bestimmt. Dabei werden sowohl inhaltliche, volitionale als auch affektive Aspekte für die Unterrichtsplanung eingesetzt.

In diesem Sinne stellt auch die Verwendung von Klassifizierungen (sonderpädagogischer Unterstützungsbedarf, Behinderung usw.) keine Problematik dar, weil diese nicht als Hinderungsgrund für die Unterrichtsgestaltung verstanden werden. Vielmehr wird Unterricht so gestaltet, dass eine maximale Partizipation aller Schülerinnen und Schüler am unterrichtlichen Geschehen ermöglicht wird. Aus einer reflexiven Perspektive ist die Verwendung von Klassifizierungen notwendig, um die Effekte der Lernumgebungen oder Assess-

Tabelle 2.1: Kriterien für einen diversitätswertschätzenden Unterricht

Kriterien nach
Sliwka (2012)
Sensibler Umgang mit menschlicher Individualität
Vorerfahrungen und Vorwissen von Lernenden ernst nehmen
Aktivierung der Lernenden
Entwicklung von einem Verständnis des eigenen Lernprozesses
Berücksichtigung von Motivation und Emotion als treibende Kräfte von Lernprozessen
Lernen als sozialen Prozess verstehen
Herstellung vielfältiger, lernförderlicher Sozialsituationen
Lernenden Herausforderungen bieten
Entwicklung in die nächste Zone ermöglichen
Transparenz in Bewertungskriterien
Leistungsrückmeldung im Dienste der Lern- und Entwicklungsförderung
Horizontale Vernetzungen zwischen Wissensgebieten und der Lebenswelt
Fischer et al. (2014)
Kooperatives Lernen
Lernen durch Lehren
Lernen am Modell durch jahrgangsübergreifendes Lernen
Individualisierte Aufgabenformate
Zeitweise Bildung von homogenen Gruppen

ments abschätzen zu können. Im Rahmen dieses Projekts werden so Stigmatisierungen durch Klassifizierungen vermieden.

2.2 Gemeinsamer Unterricht oder Lehren und Lernen für alle?

Neben den Begriffen *Heterogenität* und *Diversität* erfährt seit der Ratifizierung der UN-BRK durch die Bundesregierung 2009 der Begriff *Inklusion* verstärkte Aufmerksamkeit. *Inklusion* beschreibt oftmals den *gemeinsamen Unterricht* von Schülerinnen und Schülern mit und ohne sonderpädagogischen Unterstützungsbedarf. Diese Beschreibung ist eng mit der *negativen Perspektive* auf Lernendenmerkmale verbunden. Im internationalen Raum wird dieser Zugang als *Special-Needs Education* beschrieben und verfolgt die Idee, Adaptionen des Unterrichts für bestimmte Zielgruppen vorzunehmen. Ein solcher Unterricht problematisiert Lernendenmerkmale und stellt damit ein Ermöglichen alter Unterrichtsformen unter Adressierung neuer Gruppen dar.

Eine zweite Möglichkeit besteht in einer Orientierung an einer *Inclusive Education*, die weitere Differenzlinien miteinbezieht und mit der *positiven Perspektive* in Bezug auf Lernendenmerkmale verbunden ist. Gerade weil ein enges Inklusionsverständnis im Sinne einer *Special-Needs Education* politisch motiviert ist, wird dies oftmals abgelehnt (Ainscow, 2007; Fischer et al., 2014; Göransson & Nilholm, 2014; Walkowiak & Nehring, 2017a; Wocken, 2014). In diesem Sinne stellt der weite Inklusionsbegriff die gleiche Forderung an anspruchsvolle und erfolgreiche Lehr-Lernprozesse, die bereits schon früher bestand

(Schaefer, 1971). Geändert hat sich der Fokus auf nicht-fachliche und individuelle Voraussetzungen als Gelingensbedingungen sowie deren Interaktion für Lernprozesse (Walkowiak & Nehring, 2017a). Hiermit ist die Forderung verbunden, Unterricht als *Lehren und Lernen für alle* zu verstehen und zu konzeptualisieren.

Im Rahmen dieses Projekts wird daher der *Inklusionsbegriff* in der Konzeption von *Inclusive Education* verstanden. Letztlich stellt ein solches Verständnis eine größere Nähe zur UN-BRK sowie dem Menschenrecht auf Bildung dar (Abels, 2015). Vor diesem Hintergrund sind nur solche Unterrichtsmodelle geeignet, die verschiedenartige Differenzlinien sowie deren Dynamiken berücksichtigen. Nur dann kann im Sinne einer *Inclusive Education* von einem angemessenen Modell für einen *inklusiven Chemieunterricht* gesprochen werden (Walkowiak & Nehring, 2017a).

Aus der *negativen Perspektive* werden die *ethische und ressourcenorientierte*, die *pragmatisch-didaktische* und die *rechtlichen Ebene* von inklusiven Chemieunterricht nur implizit angesprochen (Walkowiak & Nehring, 2017a). Das ist hoch problematisch, weil damit eine nicht zielführende Vermischung von Diskussionslinien erfolgt. Aus einer *fachdidaktischen Perspektive* kann ein inklusiver naturwissenschaftlicher Unterricht nicht nur auf eine bloße soziale Teilhabe beschränkt sein (Menthe et al., 2017). Dennoch erweist sich die praktische Umsetzung als problematisch (Walkowiak & Nehring, 2017b). Ein Grund hierfür besteht in dem Problem, dass die Einführung von Inklusion auf verschiedenen Ebenen gleichzeitig erfolgte bzw. weiterhin erfolgen muss. Dazu zählen zumindest eine unterrichtspraktische, eine ethische und eine ressourcenfokussierte Ebene (Mitchel, 2014). Darüber hinaus sind die Ebenen voneinander abhängig. Entsprechend kann die Einführung von Inklusion als "eine der umfangreichsten Schulreformen der letzten 100 Jahre" beschrieben werden (Grosche, 2015, S. 18).

Die ethische und die ressourcenfokussierte Ebene

Inklusion wurde mit der Ratifizierung der UN-BRK in Deutschland bedeutsam. Für das Bildungssystem ist Artikel 24 der UN-BRK entscheidend, was die folgenden Auszüge illustrieren.

> "Bei der Verwirklichung dieses Rechts stellen die Vertragsstaaten sicher, dass Menschen mit Behinderungen nicht aufgrund ihrer Behinderung vom allgemeinen Bildungssystem ausgeschlossen werden [...]; [dass] Menschen mit Behinderungen gleichberechtigt mit anderen in der Gemeinschaft, in der sie leben, Zugang zu einem integrativen, hochwertigen und unentgeltlichen Unterricht an Grundschulen und weiterführenden Schulen haben; [dass] angemessene Vorkehrungen für die Bedürfnisse des Einzelnen getroffen werden; [dass] Menschen mit Behinderungen innerhalb des allgemeinen Bildungssystems die notwendige Unterstützung [...] [erhalten], um ihre erfolgreiche Bildung zu erleichtern" (Beauftragte der Bundesregierung für die Belange behinderter Menschen, 2014, Art. 24, 2, a–e).

Die *ethische Ebene* von Inklusion äußert sich folglich dadurch, dass einerseits eine *Gleichberechtigung* von Menschen mit unterschiedlichen Ausgangslagen hergestellt werden soll. Den Menschen soll damit Teilhabe an allen Schulformen ermöglicht werden (Walkowiak & Nehring, 2017a). Damit ist ein *normativer Anspruch* formuliert. In der Folge müssen *Bedingungen* geschaffen werden, damit alle Schülerinnen und Schüler erfolgreich Bildungsabschlüsse erwerben können, was eine *deskriptive Aussage* darstellt. Dies ist die

ressourcenorientierte Ebene. Ausgehend von dem Zitat erfährt der Begriff Inklusion durch die Kombination aus normativer und deskriptiver Aussage einen intrinsischen Wert. An diesem Punkt entwickelt sich der Konflikt zwischen der ethischen und der ressourcenorientierten Ebene von Inklusion. Der normativen Aussage stimmt vermutlich jeder zu. Bei der Frage nach den Ressourcen (deskriptive Aussage) besteht und bestand schon vor der Einführung von Inklusion ein Mangel an Ressourcen im Bildungssystem, der nun nochmals verstärkt wahrgenommen wird. Als mögliche Folge kann es zur Unterminierung des intrinsischen Werts von Inklusion kommen. Mangelnde Ressourcen lassen dann eine Absenkung der Unterrichtsqualität für alle Lernenden vermuten. Damit laufen letztlich die ressourcenorientierte und die ethische Ebene gegeneinander, weil dennoch ein normativer Anspruch besteht, einen guten Unterricht zu gewährleisten.

Die unterrichtspraktische Ebene

Ein zentrales Element des *Chemieunterrichts* stellt der authentische Einsatz von Experimenten dar. Für den *inklusiven Chemieunterricht* ergibt sich hierüber oftmals ein Konfliktpotenzial.

> "In der Wahrnehmung der Praxis erhöhen Schülerinnen und Schüler mit Förderbedarfen oder Behinderung das Gefährdungspotenzial von Experimenten. Dabei werden vor allem emotional-soziale Problematiken angesprochen, die sich in der mangelnden Disziplin der Lernenden niederschlägt, wie kognitive Schwierigkeiten als Gefahren wahrzunehmen und sicher damit umzugehen. Damit wird eine rechtliche Ebene angesprochen, die für Lehrkräfte im Extremfall nur mit dem (generellen) Ausschluss von Schülern und Schülerinnen mit Handicaps zu lösen ist" (Walkowiak & Nehring, 2017a, S. 2).

Das Experiment steht demnach exemplarisch für die Balance zwischen einer Subjekt- und einer Fachorientierung (Stroh, 2014). In diesem Sinne rekurriert das Experiment auf die Wissenschaft *Chemie* und fokussiert auf die Wissensvermittlung, während neuere Ansätze, z.B. Chemie im Kontext (Demuth, Gräsel & Parchmann, 2008), sich am Subjekt und dessen Vorstellungsentwicklung orientieren. Entsprechend wird hier das eigene Forschen im Unterrichtsgang betont. Für den Einsatz von Experimenten ist daher für jede Unterrichtssituation das Verhältnis von Fach- und Subjektorientierung neu zu bestimmen. Ein möglicher Grund für den kochbuchartigen Charakter von Experimenten liegt in der gefühlt höheren Sicherheit für Lehrkräfte, wenn sie stärker fachlich orientiert unterrichten (Hofstein & Lunetta, 2004, 1982).

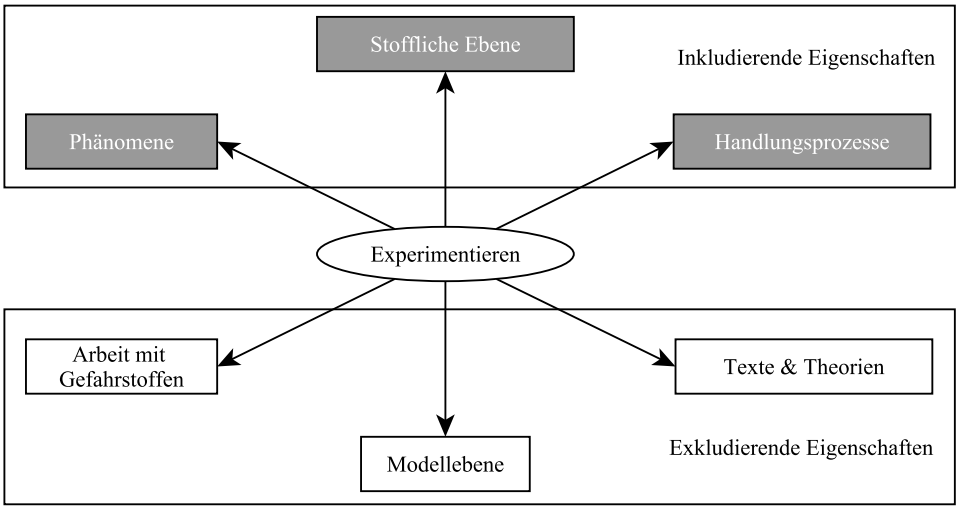


Abbildung 2.2: Exkludierende und inkludierende Aspekte des Experimentierens (nach Menthe und Hoffmann, 2015).

Um dieser Problematik gerecht zu werden, diskutieren Menthe & Hoffmann (2015) sechs mögliche Aspekte für einen inklusiven Chemieunterricht (Abb. 2.2). Den Autoren nach stellt die Phänomenorientierung (Feuererscheinungen, Farbumschläge oder farbige Niederschläge etc.) durch Experimenten verschiedenste motivierende Wahrnehmungen bereit. Werden hierüber jedoch keine fachlichen Prozesse angestoßen, stellt dies eine fachdidaktische Verkürzung dar. Begründet ist dies in der Trennung von Stoff- und Teilchenebene, die als wesentliches Strukturelement des Chemieunterrichts gelten (Johnstone, 1991). Während für nahezu alle Lernenden die stoffliche bzw. makroskopische Ebene über hör-, seh- und riechbare Beobachtungen recht unmittelbar zugänglich ist, kann dies für Erklärungen dieser Phänomene und die damit verbundene Erarbeitung von Theorien für die Teilchenebene nicht behauptet werden. Dies begründet sich vor allem im hohen Abstraktionsniveau. Da aber oftmals Modellvorstellungen für Erklärungen genutzt und darüber auch noch die Erkenntnisse in symbolische Formeln überführt werden müssen (mikroskopische bzw. symbolische Ebene), sind in Abhängigkeit zum Unterrichtsgang adäquate Repräsentationsebenen zu wählen. Gleichzeitig ermöglicht das Experimentieren einen handlungsorientierten Unterricht und bietet so vielschichtige Möglichkeiten, Handlungsprozesse anzuleiten. Hierzu zählen auch die Variationen von Arbeits- und Sozialformen, mit denen Hürden, entstanden durch den Einsatz eines Experiments, abgebaut werden können (z.B. die Anfertigung eines Laborjournals über eine ganze Reihe von Experimenten zu einer Problemstellung (Villanueva & Hand, 2011)). Letztlich birgt experimentelles Arbeiten durch den Einsatz mit Gefahrstoffen aber auch Gefahren, denen mit Regelklarheit begegnet werden muss. Wird experimentelles Arbeiten im Rahmen zur offenen Beantwortung von Problemstellungen verwendet, kann es, kombiniert mit Differenzierungsmaßnahmen, erfolgreich sein, um Inhalte zu vermitteln (Therrien, Taylor, Hosp, Kaldenberg & Gorsh, 2011; Therrien, Taylor, Watt & Kaldenberg, 2014; Villanueva, Taylor, Therrien & Hand, 2012).

Ein weiterer Punkt, der exkludierend wirken kann, besteht in der Nähe des Chemieunterrichts zum universitären Fach Chemie (Menthe & Hoffmann, 2015). Um diesen Schwierigkeiten zu begegnen, hat Mahaffy (2004) die Ebenen nach Johnstone um das *Human Element* erweitert. Dieses lebensweltliche Element soll den Chemieunterricht zugänglicher

gestalten. Dies kann unter Einbezug von sinnvollen Kontexten oder durch Schülervorstellungen geschehen. Letztlich weist der Chemieunterricht oftmals einen kanonischen Gang auf (Menthe & Hoffmann, 2015). Damit ist das gleichzeitige Nachvollziehen eines Lernschritts von allen Schülerinnen und Schülern gemeint. Auch dieser Umstand lässt sich im Verhältnis von Fach- und Subjektorientierung verorten, weil hohe Fachlichkeit mit dem Gefühl von Kontrolle und Sicherheit im Unterrichtsgang einhergeht.

2.3 Universal Design: Barrierefreiheit und Zugänglichkeit

Vor dem Hintergrund einer *Inclusive Education* ist es notwendig, jede Form von Differenzierung auch mit Blick auf den Inhalt zu wählen. Eine Möglichkeit Adaption und Verwendung gleichzeitig zu denken, besteht im Universal Design (UD). UD stellt eine Möglichkeit zur Rahmung des Entwicklungsprozesses dar, um Produkte, Geräte, Umgebungen und Systeme für möglichst viele Menschen (ohne weitere Anpassung) nutzbar zu machen. Die Konzeption von UD umfasst sieben Prinzipien, um den Entwicklungsprozess zu leiten (The Center for Universal Design, 1997).

Prinzip 1 - Breite Nutzbarkeit: Die Nutzbarkeit und Marktfähigkeit muss für diverse Nutzergruppen gegeben sein. Hier wird auf die konkrete und im besten Fall identische Verwendung durch den Benutzer abgezielt. Es geht darum, Ausgrenzung oder Stigmatisierung von Benutzern unbedingt zu vermeiden.

Prinzip 2 - Flexibilität in der Benutzung: Das Design kommt einer möglichst großen Breite von Vorlieben nach. Entsprechend werden Wahlmöglichkeiten in der Benutzung eingerichtet.

Prinzip 3 - Einfache und intuitive Benutzung: Die Benutzung des Designs ist leicht verständlich und ist u.a. unabhängig vom Wissen, der Sprachfähigkeit oder der momentanen Konzentration der Benutzerin/des Benutzers. Hierzu gilt es, Komplexität zu vermeiden und die Intuition des Benutzers/der Benutzerin zu berücksichtigen.

Prinzip 4 - Sensorisch wahrnehmbare Informationen: Das Design stellt notwendige Informationen effektiv und unabhängig von der Umgebungssituation oder den sensorischen Fähigkeiten der Benutzer/der Benutzerinnen zur Verfügung. Hierzu wird vor allem auf unterschiedliche Modi für die Präsentation von Informationen geachtet. Die Lesbarkeit von besonders wichtigen Informationen wird gewährleistet.

Prinzip 5 - Fehlertoleranz: Das Design minimiert Risiken und die negativen Konsequenzen von Fehlern oder Zufällen. Hierzu zählen die Vermeidung von Gefahrenquellen und negativen Konsequenzen.

Prinzip 6 - Niedriger körperlicher Aufwand: Das Design kann effizient und komfortabel sowie ohne körperlichen Aufwand genutzt werden.

Prinzip 7 - Größe und Platz für Zugang und Benutzung: Das Design sollte die Manipulation und die Benutzung unabhängig von den physischen Eigenschaften des Benutzers/der Benutzerin ermöglichen.

Insgesamt erweisen sich die Prinzipien jedoch als wenig trennscharf. So ließen sich die Prinzipien 4-7 auch in den Prinzipien 1-3 verorten. Jedoch vermitteln die Prinzipien eine Haltung. Damit ist die Absicht verbunden, keine Unterschiede in der Verwendung von außen erkennen zu lassen. Mit Bezug auf Gebäude schlägt sich UD im Begriff der *Barrierefreiheit* nieder. Folglich geht es um *Zugang* zu Orten bzw. deren *Zugänglichkeit*, die unter anderem über Fahrstühle, Leitstreifen oder Rampen erzeugt wird.

2.3.1 Universal Design for Learning

Aus einer *positiven Perspektive* auf Lernendenmerkmale kann nun von den Ausführungen zum UD abgeleitet werden, dass Barrieren durch den Unterricht und curriculare Vorgaben und nicht durch den Lernenden entstehen.

> "Therefore, the curriculum should be adaptable to individual differences than the other way around. In this sense, traditional curricula have the *disability*, because they only work for certain learners" (Hall, Meyer & Rose, 2012, S. 4, Herv. im Orig.).

Eine Möglichkeit UD auf die von Hall et al. (2012) beschriebenen Gedanken zu beziehen, bietet das Universal Design for Learning (UDL). Dieses wurde vom Center for Applied Special Technology (CAST) entwickelt (CAST, 2018). Ausgehend von den UD-Prinzipien stellt UDL diverse Möglichkeiten von unterrichtlichen Adaptierungsmöglichkeiten zur Verfügung. UDL folgt dabei drei Prinzipien. So stellt ein UDL-basierter Unterricht *multiple Wege der Repräsentation von Informationen* ("Was" des Lernens) bereit. Außerdem werden *multiple Wege der Verarbeitung von Informationen und der Darstellung von Lernergebnissen* ("Wie" des Lernens) angeboten. Schließlich dienen *multiple Wege zur Förderung des Lernengagements und der Lernmotivation* ("Warum" des Lernens) (CAST, 2018; Schlüter, Melle & Wember, 2016).

Die drei Prinzipien gliedern sich weiter in neun Richtlinien (Tab. 2.2). UDL verfolgt das Ziel, Lerngelegenheiten für alle Lernenden zu schaffen. Ganz im Sinne einer Inclusive Education. Hierzu sollen barrierefreie Lernumgebungen für alle Lernenden erstellt werden. Der/die Lernenden und sein/ihr Zugang zum Inhalt stehen dabei im Vordergrund.

> "UDL is an approach that seeks to address an inflexible onesizefitsall traditional curriculum that often presents barriers to struggling learners by replacing it with universally designed curriculum or curriculum that is created with the intention of including all learners" (Brownell et al., 2012, S. 81).

UDL ermöglicht aus diesem Grund einen diversitätsschätzenden Unterricht: Nicht der Lernende muss sich anpassen. Vielmehr müssen Unterrichtsinhalte ausgewählt und adaptiert werden, damit sie Barrierefreiheit aufweisen. Außerdem nimmt UDL *Differenzierungen* in Bezug auf die Tiefenstrukturen von Unterricht vor. Begründet ist dies in der Annahme, dass es ein gemeinsames Lernziel für alle Schülerinnen und Schüler geben muss (Brownell et al., 2012). Dies wird als Mindestziel angesehen. Adaptierungen jeder Art dienen dabei der Ermöglichung des Mindestziels für alle Schülerinnen und Schüler. Die Wahl der Adaptierung fällt mit Bezug zum Lernziel, dem hierzu notwendigen Lerninhalt und dessen Präsentationsforma aber auch bezüglich geeigneter Verarbeitungsprozesse und Lernprodukten aus. Letzteres kann dabei formativen und summativen Charakter aufweisen. In jedem Fall dient das Lernprodukt aber dazu, die Progression darzustellen (Hall et al.,

Tabelle 2.2: Prinzipien und Richtlinien des Universal Design for Learning (UDL) (CAST, 2018; Schlüter et al., 2016)

Multiple Unterstützungsmöglichkeiten der Repräsentation	Multiple Unterstützungsmöglichkeiten der Verarbeitung von Informationen und der Darstellung	Multiple Unterstützungsmöglichkeiten zur Förderung des Lernengagements und der Lernmotivation
1. Unterstützungsmöglichkeiten zur Perzeption des Lerninhalts	4. Verschiedene Möglichkeiten zur Interaktion mit dem Lerninhalt	7. Verschiedene Angebote zur Weckung des Lerninteresses
2. Unterstützungsmöglichkeiten zur Darstellungen von sprachlichen und symbolischen Informationen des Lerninhalts	5. Verschiedene Möglichkeiten zum Ausdruck des und zur Kommunikation über den Lerninhalt	8. Unterstützungsmöglichkeiten zur Erhaltung eines engagierten Lernens
3. Unterstützungsmöglichkeiten zum besseren Verständnis des Lerninhalts	6. Unterstützungsmöglichkeiten zur Verarbeitung des Lerninhalts	9. Unterstützungsmöglichkeiten für ein selbstreguliertes Lernen

Anmerkung:
Eine ausführliche Darstelllung findet sich unter Abschnitt 7.3.

2012; Rapp & Arndt, 2012). Nur über dieses Lernziel ergibt sich die rahmende Klammer für alle erkennbar und macht UDL umsetzbar.

Außerdem umfasst UDL Elemente von Konstruktion als auch von Instruktion, indem der Unterrichtsinhalt von der Lehrkraft aufbereitet und an die Merkmale der Lerngruppe adaptiert vorgegeben wird. Ein UDL-basierter Unterricht weist eine *Engagementphase* auf, in der ein Problemaufriss, z.B. über die Darstellung von Phänomenen oder über Vorträge gegeben wird. über die *Inputphase* wird der Lerninhalt erschlossen, um ihn in der *Outputphase* zu präsentieren. Es folgt die *Assessmentphase* mit Feedback zum Lernprodukt. Alle vier Phasen können in der konkreten Ausführung Elemente eines konstruktiven oder instruktiven Unterrichts umfassen (Rapp & Arndt, 2012).

Während UDL im internationalen Raum seit einigen Jahren Aufmerksamkeit genießt (Al-Azawei et al., 2016; Capp, 2017; Rao et al., 2014), findet es im deutschen Raum erst in den letzten Jahren Interesse. Rao et al. (2014) stellen fest, dass es in den Jahren vor Erscheinung des Artikels einen großen Teil an beschreibenden Studien gab oder diese Publikationen über die Bedeutung und Verwendung von UDL berichten. Einen Grund hierfür sehen die Autoren in der noch stattfindenden Konzeptionsphase von UDL, was nicht verwundert, weil das UDL-Modell 2011 neu konzipiert wurde. Die empirisch ausgerichteten Studien konnten jedoch nicht in allen Fällen die Frage beantworten, ob tatsächlich die UDL-basierten Interventionen zu verbesserten Lernergebnissen hinsichtlich des Inhalts führen.

Al-Azawei et al. (2016) attestieren, dass die UDL-basierte Unterrichtsgestaltung *Lern-*

barrieren - vor allem über die Bereitstellung von multiplen Repräsentationsformen für unterrichtlichen Inhalt - reduziert. Außerdem zeigen Schülerinnen und Schüler, die an UDL-basierten Unterrichten teilnahmen, eine hohe Zufriedenheit, positive Einstellung und hohes Engagement gegenüber der Instruktion. Al-Azawei et al. (2016) kritisieren jedoch, dass die Forschung zu UDL bisher auf wenige Länder mit sehr ähnlichen kulturellen und sozioökonomischen Hintergründen beschränkt ist. Es ist daher fraglich, ob bestimmte Kombinationen von Lernendenmerkmalen die Wirkungen von UDL ändern. Entsprechend müssen die Ergebnisse vorsichtig auf neue regionale Situationen und Zustände übertragen werden. Schließlich bemängeln Al-Azawei et al. (2016) auch, dass es kaum UDL-Ansätze im Feld gibt, die alle drei Prinzipien umsetzen. Die referierten Studien fokussieren auf einzelne Prinzipien, wobei das Prinzip der multiplen Repräsentation die größte Aufmerksamkeit genießt.

Die Metaanalyse von Capp (2017) zeigt ebenfalls, dass bis heute das UDL-Forschungsfeld vor allem auf das Prinzip der multiplen Repräsentationen fokussiert. Das ist problematisch, weil UDL zwar nicht zwangsläufig die Umsetzung aller Prinzipien, Richtlinien und Checkpoints fordert - falls dies mit Bezug auf die Checkpoints überhaupt möglich ist. Jedoch will sich UDL als ganzheitliches Modell zur Unterrichtsplanung verstanden wissen. Außerdem gibt es wenige Berichte über Subgruppen und wie sich diese in UDL-Interventionen verhalten haben. Bezüglich der Befunde unterscheidet Capp (2017) zwischen *primären und sekundären Bildungsresultaten*. Als problematisch erweisen sich die primären Bildungszuwächse.

> "The learning outcomes associated with the implementation of UDL need be demonstrated through experimental studies within curriculum areas" (Capp, 2017, S. 804).

Insgesamt werden bisher oftmals qualitative Studien verwendet, um UDL-Implementationen zu beforschen. Mit diesem Zugang kann jedoch nicht die Frage beantwortet werden, ob es möglich ist, mit UDL-Lernumgebungen die Bedeutung von Lernendenmerkmalen (z.B. der Lesefähigkeit oder des Migrationshintergrunds) für *den Lernprozess und das Lernresultat* abzuschwächen oder zu egalisieren (Autorengruppe Bildungsberichterstattung, 2016, 2018). Hierzu sind (quasi-)experimentelle Studien notwendig, die den Einfluss von Lernendenmerkmalen auf primäre Bildungszuwächse bestimmen. Zweitens besteht ein Mangel an Studien, die curriculare Inhalte adressieren; eben jene primären Bildungsresultate. Allerdings gibt es durch die Metaanalyse Evidenz dafür, dass sekundäre Bildungsresultate durch UDL positiv beeinflusst werden.

> The results of this meta-analysis support the claims made by the Center for Applied Special Technology regarding the effectiveness of UDL in improving the learning process for all students" (Capp, 2017, S. 805).

Ein Grund für diese Situation könnte darin bestehen, dass bisher weniger über das Assessment der Studien berichtet wurde und unter Umständen diesem weniger Aufmerksamkeit gewidmet wurde als den Lernumgebungen selber. Das Assessment ist jedoch von größter Bedeutung, weil die Daten gerade über selbiges erzeugt werden. Umfasst dieses auch Barrieren, kann die Wirkung der Intervention mitunter nicht adäquat abgebildet werden.

Bisweilen ist UDL auch nicht ohne Kritik geblieben. Für Edyburn (2010) bestehen zwei grundsätzliche Probleme an UDL. Zuerst gibt es Zweifel an der übertragbarkeit eines markt- bzw. architekturorientierten Konzeptes auf Bildungsprozesse.

> "For example, the interactions between individuals and the built environment (e.g. stairs, doorways, countertops) are static and limited" (Edyburn, 2010, S. 36).

Ein zweites Problem an UDL besteht für Edyburn (2010) in der Frage nach der Wertschätzungsorientierung: Wird Diversität oder Technologie wertgeschätzt? Der Erfolg von UDL hängt für den Autor mit der Verfügbarkeit von *neuen Medien* zusammen. Diese erweisen sich als variabel und bieten diverse Unterstützungsmöglichkeiten. Es besteht daher die Gefahr, dass UDL mit dem Einsatz von *neuen Medien* gleichgesetzt wird.

> "I have often observed situations where teachers, administrators, and publishers claim they are implementing UDL simply because they are using multimedia or Web 2.0 tools" (Edyburn, 2010, S. 36).

Viel entscheidender ist jedoch, dass *Vielfalt wertgeschätzt* wird. Hieraus ergibt sich eine Haltung, die proaktiv unterrichtliche Unterstützungen entwickelt. Andernfalls ist das Ergebnis in Form einer Lernumgebung einfach ein glücklicher Zufall zwischen dem Einsatz von neuen Medien und dem *Lerninhalt*. UDL ist und soll weit mehr als nur die Integration der neuesten Technologien in Lehrpläne darstellen. Gleichzeitig ist aber Technologie für die Implementierung von UDL unerlässlich, gerade weil sie Potenziale für eine adaptive Nutzung hat (z.B. Vorlesefunktionen oder Bildschirmeinstellungen). Hiermit eng verbunden ist die Frage nach UDL-Implementation. UDL ist eine Form des adaptiven Unterrichts mit besonderem Fokus auf den Inhalt. Vor diesem Hintergrund wird deutlich, dass UDL nicht natürlich auftritt. Es bedarf der Aus- und Weiterbildung von Lehrerinnen und Lehrern. Es müssen sowohl die Umsetzungstechniken erlernt als auch die notwendige Einstellung zur Diversität der Lernenden ausgebildet werden. Edyburn (2010) formuliert jedoch Zweifel an der überprüfbarkeit von UDL-Implementierungen.

> "All three of the 'multiple means' statements by CAST focus on providing multiple concurrent interventions. As a result, within existing conceptualizations of UDL, there is no clear way to measure claims that UDL is effective for enhancing the academic performance of diverse students. This is a significant shortcoming for anyone trying to operationalize, implement, and evaluate a UDL program" (Edyburn, 2010, S. 39, Herv. im Orig.)

Die Problematik besteht genau in den multiplen Möglichkeiten der Konzeptualisierung von UDL. Wird diese nicht genau genug vorgenommen, können Adaptionen gegeneinander wirken. Interventionen müssen diesen Punkt berücksichtigen und nicht wahllos möglichst viele Elemente umsetzen. Gleichwohl sind auch die Elemente nicht immer eindeutig einzelnen Adaptierungen zu zuordnen und umgekehrt. Für Edyburn (2010) ist es daher zweifelsohne notwendig, die primären und sekundären Bildungsresultate von UDL zu messen.

2.3.2 Universal Design for Assessment

Aus Capp (2017) und Edyburn (2010) kann abgeleitet werden, dass dem Assessment in der Beforschung von UDL-Implementationen mehr Aufmerksamkeit zukommen sollten. Eine Möglichkeit das Assessment mit Blick auf Zugänglichkeit und Barrierefreiheit zu betrachten, besteht im Universal Design for Assessment (UDA) (Beddow, 2011; Lovett & Lewandowski, 2015; Thompson, Thurlow & Malouf, 2004). Hierzu gehört die Diskussion

und Implementierung von Testadaptionen aber auch die Wahl geeigneter statistischer Verfahren.

Das Gerechtigkeitsargument beim Einsatz von Assessments

Weit weniger prominent als das UDL ist das UDA. Das mag unter anderem auch mit der Frage nach der Fairness in Assessments einhergehen. Das *Gerechtigkeitsargument* basiert auf einem anerkannten Leistungsprinzip: Jede/r Lernende kann eine bestmögliche Leistung erbringen, wenn sie/er nur hart genug arbeitet.

Folgt man dieser Aussage, wäre es nur logisch zu sagen, dass Testadaptionen in jeder Form zurückzuweisen sind. Weiter wäre aus dieser Perspektive jede Testung genau dann gerecht, wenn jeder und jede den gleichen Test bekommt. Das *Gerechtigkeitsargument* erscheint jedoch auf den zweiten Blick und vor dem Hintergrund von Forschungsbefunden rund um die Bildungswissenschaften zweifelhaft. Zumindest im Bereich der naturwissenschaftlichen Bildung hängen die Lernergebnisse mit diversen Lernendenmerkmalen der Schülerinnen und Schüler zusammen. Hierzu zählen neben dem Geschlecht (Milgram, 2011) auch die Lesefähigkeit (Patterson, Roman, Friend, Osborne & Donovan, 2018), der sozioökonomische Status, die Intelligenz sowie die Motivation (Nehring, Nowak, Upmeier & Tiemann, 2015). Letztlich erweist sich das *Gerechtigkeitsargument* auf einer dritten Ebene, nämlich der von politischen Entscheidungen, als unhaltbar. Bildungserfolg wird über *Leistungsmessungen* bestimmt. Nach der UN-BRK muss aber auch die Leistungssituation adaptiv gestaltet werden. Vor diesem Hintergrund ergeben sich auch Probleme hinsichtlich der *Reliabilität* und *Validität* von Tests, wenn diese *in diversen Lerngruppen* eingesetzt werden, aber nicht auf *Diversität* ausgerichtet sind.

Zu dieser Problematik hat der National Research Council (1998) (NRC) Standards für den Entwicklungsprozess von Tests formuliert. Demnach ergibt sich *Gerechtigkeit als Testeigenschaft*, wie Validität, nicht als ein nachträglicher Zufall, sondern muss von Beginn an betrachtet werden. Es muss von der Testgestaltung und -entwicklung bis hin zur Bewertung, Interpretation und Nutzung der Daten bedacht werden (National Research Council, 1998). UDA bietet hierfür einen theoretischen Rahmen, um mit den dargestellten Anforderungen umzugehen.

> " 'Universally designed assessments' are designed and developed from the beginning to allow participation of the widest possible range of students, and to result in valid inferences about performance for all students who participate in the assessment. Universally designed assessments add a dimension of fairness to the testing process" (Thompson et al., 2002, S. 5, Herv. im Orig.).

Das UDA basiert auf der Prämisse, dass Testteilnehmende bestmögliche Testergebnisse, unabhängig von seinen/ihren Personenmerkmalen erreicht. Zum Erstellen von UDA-basierten Tests werden ähnlich wie beim UDL Elemente formuliert (Tab. 2.3). Für die vorliegende Arbeit ist es daher notwendig, die Evaluation von Bildungsprozessen nicht ein- sondern mehrdimensional zu betrachten (Gesellschaft für Fachdidaktik, 2017; Stellungnahme der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), 2017). In der Folge sollten nicht einzelne Differenz- und Ungleichheitsdimensionen, sondern deren Interaktion berücksichtigt werden. Entsprechend fand die Auswahl der Lernendenmerkmale in dieser Studie auf Basis aktueller Bildungsberichterstattung statt (Autorengruppe Bildungsberichterstattung, 2016, 2018).

Die Erhöhung der Testzugänglichkeit

Tabelle 2.3: Elemente für einen universell designten Test (Thompson und Thurlow, 2004)

Element	Beschreibung
1. Inclusive assessment population	Test findet in diversen Lerngruppen statt. Nur so können generalisierbare Aussagen abgeleitet werden.
2. Precisely defined constructs	Das Konstrukt des Tests ist theoretisch sehr gut definiert. Die Wahl der Testadaptionen ist hieraufhin zu überprüfen.
3. Accessible, non-biased items	Die Beantwortung der Items ist unabhängig von den Lernendenmerkmalen.
4. Amenable to accommodations	Die Testadaptionen sind durch die Testteilnehmenden veränderbar.
5. Simple, clear, and intuitive instructions and procedures	Die Testdurchführung muss klar und verständlich sein.
6. Maximum readability and comprehensibility	Textformatierungen müssen gut lesbar sein.
7. Maximum legibility	Der Test muss für die Testteilnehmenden verständlich sein.

Die UDA-Elemente (Tab. 2.3) können dabei auf zwei Weisen dienlich sein: Zuerst geben sie Anhaltspunkte für die theoretischen Aspekte der Testentwicklung. UDA-basierte Assessments versuchen die Barrierefreiheit zu verbessern bzw. herzustellen, indem *Testbarrieren* abgebaut werden. In diesem Zusammenhang steht der Begriff *Testzugänglichkeit*. Dieser basiert auf der Idee, dass Elemente eines Tests es dem Testteilnehmer/der Testteilnehmerin ermöglichen, ihr/sein Wissen oder seine/ihre Fähigkeiten relativ zum Zielkonstrukt und ohne Ablenkung oder Verzerrung zu demonstrieren. Der *Testzugang* wird daher definiert als die Interaktion zwischen konstruktirrelevanten Item- und Personenmerkmalen (Winter, Kopriva, Chen & Emick, 2006).

Dementsprechend können drei Testzugänglichkeitshinweise formuliert werden (Abb. 2.3). Erstens sollen die Testteilnehmenden nur mit dem Zielkonstrukt des Tests interagieren. In der Folge werden nur intrinsische Eigenschaften der Testteilnehmenden in Bezug auf das Zielkonstrukt des Tests gemessen. Zweitens wird die Barrierefreiheit dort etabliert, wo Testfunktionen mit diesen individuellen Eigenschaften interagieren. Drittens können Testadaptionen in Kombination verwendet werden, um den Einfluss von Nebeninteraktionen zu reduzieren und den Zugang für Testteilnehmende zu erhöhen (Beddow, 2011).

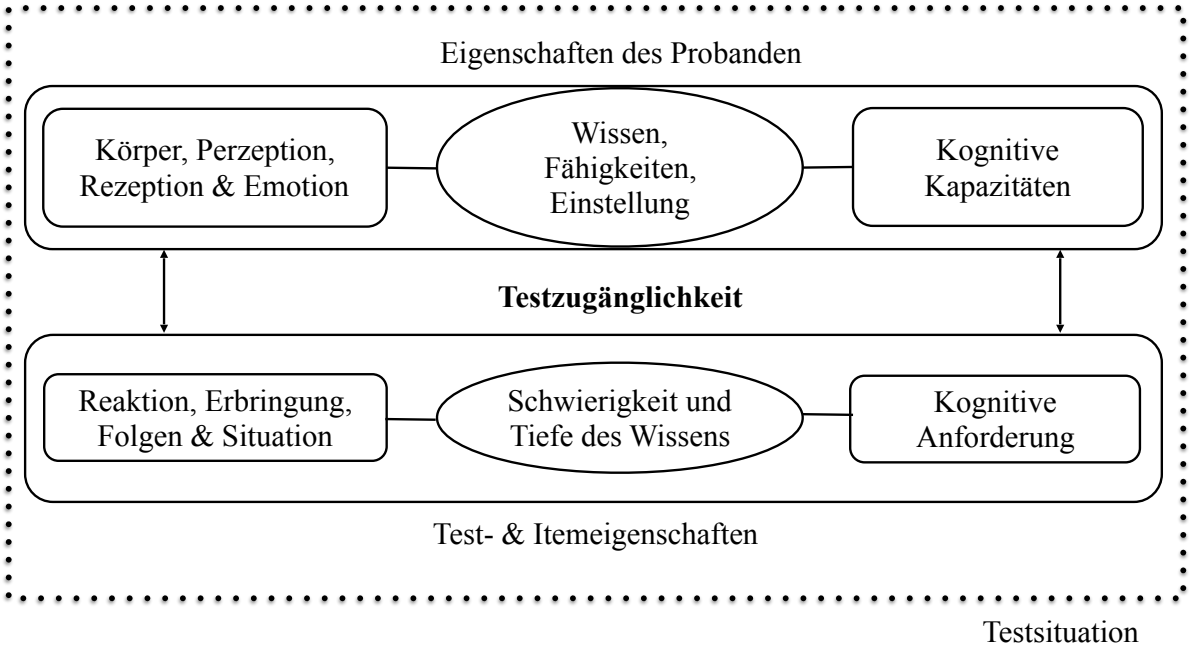


Abbildung 2.3: Modell für die Prüfung von Testzugänglichkeit (Beddow, 2011, 388, übersetzung d. Verf.).

Entscheidend ist dabei die Einsicht, dass nur wenn das Zielkonstrukt eindeutig gemessen wird, der Testwert überhaupt einen Rückschluss zum Wissen, der Fähigkeit oder der Einstellung zulässt. *Testzugänglichkeit* steht somit eng in Verbindung mit der Reduzierung von *konstruktirrelevanter Varianz* (KIV) (Messick, 1984). Eine Möglichkeit die KIV inhaltlich zu greifen, besteht im Gedanken des systematischen Messfehlers.

> "Systematic error is not random, but group- or personspecific. Construct-irrelevant easiness refers to a contaminating influence on test scores that tends to systematically increase test scores for a specific examinee or a group of examinees; construct-irrelevant difficulty does the opposite. It systematically decreases test scores for a specific examinee or a group of examinees" (Haladyna, State & Downing, 2004, S. 18).

Daraus lässt sich ableiten, dass nur, wenn die KIV gering ist, der Testwert eines Probanden valide und reliabel ist. Erst wenn diese Annahme zutrifft, lassen sich Rückschlüsse über das Zielkonstrukt ziehen.

Die Literatur um UDA stellt neben den Elementen auch Möglichkeiten für eine geeignete statistische Methode zur Bewertung der Daten bereit (Johnstone, 2003; Thompson et al., 2002). So fokussieren frühe überlegungen vor allem auf das UDA-Element "Accessible, non-biased items" und schlagen Analysen zum *Differential Item Functioning* (DIF) vor, um die Testzugänglichkeit zu prüfen. Hierbei wird die Gruppenabhängigkeit von Items

untersucht, das heißt, dass nach unterschiedlichen Antwortverhalten von Gruppen zum gleichen Item gesucht wird. Hierzu werden Subgruppenanalysen auf Basis der erhobenen Lernendenmerkmalen vorgenommen. Lovett & Lewandowski (2015) schlägt darüber hinaus vor, die Vergleichbarkeit des Konstrukts zu prüfen, wenn ein Test mit und ohne Testadaptionen verwendet wurde. Hierbei wird auf das UDA-Element "Precisely defined constructs" verwiesen. Hierfür werden sowohl *explorative und konfirmatorische Faktorenanalysen* vorgeschlagen.

Der Differential Boost

Ein Grundgedanke in der Diskussion um Testadaptionen besteht darin, dass einige Testpersonen diese benötigen, andere jedoch nicht. Es geht aber auch weiter darum, die angenommene, durch Adaptionen ermöglichte Fairness, zu überprüfen. Wenn jedoch die Testadaptionen zu einem unfairen Vorteil führen sollten, dann könnten die Daten einem Bias unterliegen, was für Testteilnehmende mit nicht adaptierten Tests einen Nachteil darstellen würde. Aus diesem Grund wurde der *Differential Boost* als Beobachtungskriterium für die *Interaction Hypothesis* vorgeschlagen. Die *Interaction Hypothesis* dient der Rechtfertigung von Testadaptionen (Elliott, Kettler, Beddow & Kurz, 2018; Koenig & Bachman, 2004; Lovett & Lewandowski, 2015; Shepard, Taylor & Betebenner, 1998; Sireci, Scarpati & Li, 2005; Weston, 2002). Die *Interaction Hypothesis* besagt einerseits, dass eingesetzte Testadaptionen zu besseren Testergebnissen bei Schülerinnen und Schülern mit Behinderungen führen als unter Standardbedingungen. Andererseits, dass Schülerinnen und Schüler ohne Behinderungen keine besseren Testergebnisse aufweisen, wenn sie adaptierte Testinstrumente nutzen. Oftmals wird die Wirkung von Testadaptionen über die Bestimmung von Effektstärken vorgenommen. Ist diese bei der Gruppe mit Schülerinnen und Schülern mit Behinderungen größer als bei der Vergleichsgruppe, liegt ein *Differential Boost* vor.

Obwohl der Gedanke naheliegt, dass KIV nicht nur bei Schülerinnen und Schülern mit Behinderungen, sondern in einer ganzen Reihe von Situationen auftreten kann, findet erst allmählich eine öffnung der Frage nach der *Interaction Hypothesis* mit Blick auf andere Differenzlinien statt (Sireci, Banda & Wells, 2018; Witmer, Schmitt, Clinton & Mathes, 2018).

Fuchs, Fuchs & Capizzi (2005) und Sireci et al. (2005) untersuchten in Meta-Studien Publikationen, die der Frage nach einem *Differential Boost* nachgehen. Die Testadaptionen in diesen Studien umfassen das Vorlesen, zusätzliche Testzeit und weitere, kombinierte Testadaptionen. Auch das National Center on Education Outcomes (NCEO) in den USA veröffentlicht regelmäßig eine übersicht zu Untersuchungen zur *Interaction Hypothesis* (Cormier, Altman, Shyyan & Thurlow, 2010; Thurlow, Christian & Rogers, 2012; Thurlow, Lazarus & Rogers, 2014). über alle Publikationen hinweg zeichnet sich ein gemischtes Bild zur Frage nach einem *Differential Boost* ab. Oft wurde berichtet, dass kein *Differential Boost* beobachtet werden konnte. Dies gilt vor allem für Studien, die die Testzeit für die Gruppe von Schülerinnen und Schüler mit Behinderungen verlängert haben. Studien mit computerbasierten Tests und mit Vorlesefunktionen konnten dagegen einen *Differential Boost* beobachten.

In den Studien, die keinen *Differential Boost* beobachteten, wurden alternative Erklärungen angeführt. Diese bezogen sich im Allgemeinen auf andere Formen von Validitätsnachweisen zur überprüfung der Eignung von Testadaptionen. Aufgrund dieser uneindeutigen Befundlage werden auch andere überprüfungsformen empfohlen. So werden nun Nachwei-

se über die Reliabilität der Testergebnisse eingefordert (AERA, APA & NCEO, 2014). Die meisten Untersuchungen zu Testadaptionen haben sich jedoch bisher nicht auf die Reliabilität konzentriert - stattdessen wurden Mittelwertsunterschiede und Effektstärken fokussiert. Daher gibt es keine etablierten Erwartungen für die Auswirkungen von Testadaptionen auf Reliabilitätsindizes (Elliott et al., 2018). Entsprechend werden interne Konsistenzen für einzelne Testteile oder den Test bestimmt. Problematisch bei der Verwendung von internen Konsistenzschätzungen ist, dass die Repräsentation für die Zuverlässigkeit darin besteht, dass Korrelationen zwischen Teilmengen von Items und nicht zwischen dem Gesamtergebnis des Testergebnisses bestimmt werden. Letzteres wird aber oftmals zur Testinterpretation verwendetet (Elliott et al., 2018). Testadaptionen sollten die interne Konsistenz von Skalen nicht verringern. Im Gegenteil sollte sich die interne Konsistenz einer Gruppe von Schülerinnen und Schülern mit Risikofaktoren durch Testadaptionen erhöhen (Elliott et al., 2018).

2.4 Die Nature of Science (NOS): Eine Reflexionspraxis

Einen möglichen Anknüpfungspunkt zwischen der *Bedeutung des Experiments* und der *Ausbildung einer naturwissenschaftlichen Grundbildung* durch den naturwissenschaftlichen Unterricht bietet die *Nature of Science* (NOS) (Abd-El-Khalick & Lederman, 2000; Holbrook & Rannikmae, 2007; Lederman, 2013). NOS spielt eine Schlüsselrolle in curricularen Standards und ist seit vielen Jahren Gegenstand nationaler und internationaler fachdidaktischer Forschung (Bernholt et al., 2012; NGSS Lead States, 2013).

Driver, Leach & Millar (1996) führen die Bedeutung von NOS an fünf möglichen Vorteilen an.

1. NOS ist demnach bedeutsam, um den Prozess der Wissenschaft verstehen zu können.
2. Außerdem hilft NOS, fundierte Entscheidungen zu naturwissenschaftlichen Themen treffen zu können.
3. Darüber hinaus stellen die Naturwissenschaften ein zentrales Element der zeitgenössischen Kultur dar.
4. Schließlich umfasst NOS Normen des Wissenschaftlichen und hilft, sich dieser stärker bewusst zu sein.
5. Letztlich erfährt das Lernen naturwissenschaftlicher Inhalte mehr Tiefe, wenn Aspekte von NOS mit einfließen.

NOS schlägt damit eine Brücke zwischen dem Bild und dem Fachwissen der Naturwissenschaften bei den Schülerinnen und Schülern sowie der Arbeit von Naturwissenschaftlern und der Bedeutung der Naturwissenschaften für Gesellschaften. Exemplarisch sei auf die Socio-Scientific-Issues (SSI) verwiesen, welche gesellschaftlich relevante Probleme thematisieren (z.B. Einlagerung von Atommüllfässern in den Salzstock Gorleben). Diese Probleme sind allerdings nicht nur auf naturwissenschaftlicher Basis schwierig zu entscheiden, sondern können auch politische, soziale oder ethische Aspekte umfassen. über NOS lässt sich daher aus fachdidaktischer Perspektive eine Verbindung des Unterrichts in die Lebenswelt und die Biografie der Lernenden herstellen. Leider verfügen Schülerinnen und Schüler aber nur über wenig elaborierte NOS-Konzepte.

> "Die vorliegenden empirischen Studien zum Verständnis von Kindern und Jugendlichen über die Natur der Naturwissenschaften zeigen kein einheitliches Bild. [...] Zusammenfassend lässt sich sagen, dass die Vorstellungen, die sich Schüler und Schülerinnen von der Natur der Naturwissenschaften machen, als unzureichend und nicht adäquat bezeichnet werden müssen. Unter einem Naturwissenschaftler stellen sie sich tendenziell ein männliches und absonderliches Stereotyp vor. Er ist mal gefährlich, mal wissensdurstig, mal hilfreich, aber immer ist er fremd. Als primäre Motivation scheint ein diffuser Wissensdrang entscheidend zu sein. Die Vorstellungen zum epistemologischen Status naturwissenschaftlicher Wissensbestände zeigt eine Tendenz zum ontologischen Realismus" (Höttecke, 2001, S. 20).

Neben der Frage nach den Eigenschaften der Naturwissenschaften und dem naturwissenschaftlichen Wissen, besteht auch die Frage nach den Arbeitsweisen von Naturwissenschaftlern und Naturwissenschaftlerinnen selbst. Höttecke (2001) beschreibt diese als "naivempiristisch". Das Experimentieren wird mit der Sammlung von Beobachtungsgaben gleichgesetzt. Hinzu kommt gerade bei jüngeren Schülerinnen und Schülern die Vorstellung eines planlosen Ausprobierens und Entdeckens im Experimentierprozess (Carey, Evans, Honda, Jay & Unger, 1989). Auch sehen sich Schülerinnen und Schüler nicht dazu befähigt, naturwissenschaftliches Wissen zu erarbeiten. Die kulturelle und soziale Einbettung von Naturwissenschaften wird nicht erkannt. Entsprechend besteht die Vorstellung, dass Naturwissenschaftlerinnen und Naturwissenschaftler einzeln und isoliert arbeiten.

> "Dieser Umstand ist im Sinne einer naturwissenschaftlichen Grund- und Allgemeinbildung nicht akzeptabel. Daher ergibt sich ein starker Aufforderungsimpuls an den naturwissenschaftlichen Unterricht und die jeweiligen Fachdidaktiken, einen Beitrag zur Verbesserung der Situation zu leisten" (Höttecke, 2001, S. 21).

2.4.1 Die Inhaltsbereiche von NOS und ihre ähnlichkeit zu Epistimic Beliefs

Um die Inhaltsbereiche von NOS wurden immer Bemühungen über verschiedenste Methoden zur Erstellung von Konsenslisten betrieben (Delphi-Studie: Osborne, Collins, Ratcliffe, Millar & Duschl (2003); Analyse von nationalen Standards und Lehrplänen: McComas & Olson (1998)). Nach Lederman (1992) bezieht sich NOS auf die Erkenntnistheorie der Naturwissenschaften oder als eine Sammlung von Werten und Annahmen. All dies ist Teil der naturwissenschaftlichen Erkenntnisgewinnung und des naturwissenschaftlichen Wissens. Ausgehend von diesem Punkt umfasst NOS u.a. Konzepte zur Theoretisierung von naturwissenschaftlichen Wissen, dessen empirischer Natur oder aber auch zum Unterschied von Theorien und Gesetzen oder der sozialen und kulturellen Einbettung naturwissenschaftlichen Wissens (Lederman, 2013; Lederman, Abd-El-Khalick, Bell & Schwartz, 2002). Problematisch an Konsenslisten zur Bestimmung der NOS-Inhaltsbereiche ist jedoch ihr statischer Charakter. Abd-El-Khalick & Lederman (2000) argumentieren daher, dass Veränderungen in der Auffassung, was die Naturwissenschaft auszeichnet, notwendigerweise die Frage nach den Inhaltsbereichen von NOS neu aufwirft.

Neben dieser eher wissenschaftstheoretischen Perspektive gibt es auch einen eher kognitionspsychlogischen Ansatz, der auf die Konzeptualisierung von *Epistimic Beliefs* (EBs)

zielt (Neumann & Kremer, 2013). In diesen Feld gibt es zwei Hauptfragen: 1) Was ändert sich bezüglich der EBs?; 2) Wie kann die Art der Veränderungen der EB beschrieben werden? (Conley, Pintrich, Vekiri & Harrison, 2004). Bezüglich der ersten Frage besteht eine ähnlichkeit zum NOS-Feld in der andauernden Debatte um relevante Inhaltsbereiche (Conley et al., 2004; Pintrich, 2002). Ein Fixpunkt für beide Forschungsfelder stellt jedoch die Annahme dar, dass es EBs/Konzepte zur Natur des Wissens und dessen Nutzung geben muss (Conley et al., 2004; Hofer & Pintrich, 1997). Conley et al. (2004) entwickelten auf Basis einer umfangreichen Literaturrecherche ein vier dimensionales Model zu EBs. Dieses Modell umfasst EBs zur Sicherheit (certainty), der Herkunft (source), der Entwicklung (development) und der Rechtfertigung (justification) von Wissen (Elder, 2002; Hofer, 2000; Schommer, 1990; Schraw, Bendixen & Dunkle, 2002).

Die Dimensionen *Herkunft* und *Rechtfertigung* reflektieren EBs über die *Natur der Wissenskonstruktion*. Weniger elaborierte EBs der Dimension *Herkunft* verstehen Wissen nicht als konstruiert, sondern externen Autoritäten innewohnend. Das Selbst ist entsprechend nicht Konstrukteur des Wissens. Die Dimension *Rechtfertigung* thematisiert die Verwendung von Daten und der Methoden zur Datengewinnung (z.B.: über Experimente) zur Unterstützung von Argumenten. Die anderen beiden Dimensionen spiegeln EBs über die *Natur des Wissens* wider. Weniger elaborierte EBs der Dimension *Sicherheit* spiegeln den Glauben an *eine* richtige Antwort wider. Im Gegensatz dazu führen elaborierte EBSs zu Annahmen, dass komplexe Probleme auf verschiedene Weisen gelöst werden können. Die Dimension *Entwicklung* umfasst EBs, die Wissenschaft als sich entwickelndes Konstrukt anerkennt. Demnach können sich Ideen, Theorien und Gesetzmäßigkeiten auf der Grundlage neuer Daten und Erkenntnisse verändern. (Conley et al., 2004)

Die ähnlichkeiten und Unterschiede der Forschungstraditionen von NOS und EBs sind unverkennbar und vor allem vor dem Hintergrund der Forschungstraditionen zu erkennen (Neumann & Kremer, 2013). Dies zeigt sich beispielsweise daran, dass die Inhaltsbereiche von NOS vor allem der Förderung einer naturwissenschaftlichen Grundbildung dienen (Bernholt et al., 2012; Holbrook & Rannikmae, 2007; Lederman, 2013; Neumann & Kremer, 2013; NGSS Lead States, 2013). Entsprechend können Testinventare zu NOS-Konzepten entwickelt werden, wenn diese die *naturwissenschaftliche Wissensproduktion* und die *Eigenschaften des naturwissenschaftlichen Wissens* umfassen (Neumann & Kremer, 2013; Urhahne et al., 2008).

Kampa et al. (2016) haben ein Testinventar zu NOS-Inhaltsbereichen verwendet (Urhahne et al., 2008) und mit Bezug zu Conley et al. (2004) adaptiert (Tab. 2.4). Das Testinventar umfasst negativ formulierte Items zur *Sicherheit des naturwissenschaftlichen Wissens*, was zwar relativ verlässlich und dauerhaft sein kann, aber stets Vorläufigkeitscharakter aufweist. Demnach können auch verschiedene Theorien ein und dasselbe Phänomen erklären. Die EBs dieser Dimension beschreiben, dass Wissen entweder richtig oder falsch ist und gehen weiter dazu über, dass naturwissenschaftliches Wissen als Ergebnis einer multiperspektivischen Reflexion zu sehen ist. Die Dimension *Entwicklung von naturwissenschaftlichem Wissen* beschreibt die fortwährende Entwicklung des naturwissenschaftlichen Wissens. So werden einerseits EBs adressiert, die das naturwissenschaftliche Wissen als statisch und unveränderlich beschreiben. Andererseits aber, dass naturwissenschaftliche Ideen und Theorien sich im Laufe der Zeit und durch neue Evidenzen ändern. Die EBs der *Rechtfertigung zum naturwissenschaftlichen Wissen* beziehen sich vor allem auf die evidenz- und theoriebasierte Argumentation um Experimente. Diese Dimension umfasst die Bedeutung von Beobachtungen und Experimenten sowie die Begründungen zur

Tabelle 2.4: Beispielitemformulierungen von Items aus dem Testinventar von Kampa et al. (2016)

Dimension	Formulierung		Itemanzahl
Herkunft	Nur Naturwissenschaftler können die Natur beobachten.	(-)	5
Sicherheit	Das Wissen in den Naturwissenschaften ist für alle Zeiten wahr.	(-)	7
Entwicklung	Manchmal ändert sich das Wissen in den Naturwissenschaften.	(+)	8
Rechtfertigung	Gute Theorien stützen sich auf Ergebnisse aus vielen verschiedenen Experimenten.	(+)	7

Anmerkung:
(-): negative Formulierung
(+): positive Formulierung

Generierung von naturwissenschaftlichem Wissen. EBs dieser Dimension umfassen die Entdeckung und Theoretisierung von Phänomenen durch naturwissenschaftliche Untersuchungen (z.B. über das Experiment oder die Beobachtung). Darüber hinaus werden EBs adressiert, die ein Verständnis dafür aufzeigen, dass Wissen dem eigenen Denken und auf multiplen Experimenten und Beobachtungen basiert. Den letzten Inhaltsbereich stellt die *Herkunft von naturwissenschaftlichem Wissen* dar. So kann naturwissenschaftliches Wissen nicht bloßvon Naturwissenschaftlerinnen und Naturwissenschaftlern gebildet werden, sondern auch von Schülerinnen und Schülern, indem sie Neugierde zeigen.

2.4.2 Die Vermittlungsarten von NOS und der hierzu notwendige Grad der Kontextualisierung

Die Fragen nach der Art und Weise der NOS-Vermittlung sowie die dazugehörige Tätigkeit und ihr Kontext bestimmen bis heute das Forschungsfeld. Khishfe & Abd-El-Khalick (2002) fassen die Versuche zur Vermittlung in drei Ansätzen zusammen. Der *historische Ansatz* schlägt vor, dass die Geschichte der Naturwissenschaft in den naturwissenschaftlichen Unterricht integriert wird. Befunde für die Wirksamkeit des *historischen Ansatzes* sind jedoch bestenfalls uneindeutig (Khishfe & Abd-El-Khalick, 2002). Der *implizite Ansatz* geht davon aus, dass die Schülerinnen und Schüler durch das Praktizieren von Naturwissenschaften ihre NOS-Konzepte elaborieren (Lawson, 1982; Rowe, 1974). Dieser Ansatz sieht keine Verwendung von Verweisen auf NOS-Aspekte vor, sondern schlägt forschungsorientierte Lernaktivitäten vor (Crumb, 1965). Der *explizit-reflektierende* Ansatz nutzt vor allem dekontextualisierte Tätigkeiten zur NOS-Vermittlung über die anschließend reflektiert wird. So werden beispielsweise Black-Box-Modelle verwendet, um Methoden zur Erschließung selbiger einzusetzen und über diese zu diskutieren. Das Ziel besteht nicht

darin Wissen über die innere Struktur des zu modellierenden Systems zu gewinnen, sondern über den Prozess der Wissensproduktion (Lederman & AbdElKhalick, 1998). Vor allem mit dem *explizit-reflektierenden* Ansatz konnten wirksame Interventionen durchgeführt werden (Akerson & Hanuscin, 2007; Mulvey, Chiu, Ghosh & Bell, 2016; Schwartz, Lederman & Crawford, 2004).

In der Folge der Frage nach dem Vermittlungsansatz von NOS wird durch Clough (2006) die Frage nach dem Kontext der Vermittlung angestoßen. Ein Hauptunterschied zwischen den verschiedenen Studien besteht in der Art und dem Grad der Kontextualisierung der NOS-Vermittlung. Ein Problem hierin könnte dahingehend bestehen, dass über dekontextualisierte Tätigkeiten zwar NOS-Konzepte entwickelt, diese aber nicht angewendet werden können.

> "So while explicit/reflective decontextualized NOS teaching is important for drawing students' attention to particular NOS issues and serving as analogies to authentic science, it alone is likely insufficient for developing in students and teachers a deep understanding of the NOS that can be robustly applied in differing content-specific situations" (Clough, 2006, S. 474).

Mulvey et al. (2016) empfehlen ausgehend von ihrer Studie eine allmähliche Zunahme der Kontextualisierung.

> "Before discussing important theories and laws in science, helping students to know what the terms mean and how they are used by scientists can elevate the quality of discussions about specific theories and laws. For example, the idea that scientific theories are only hypotheses yet to be proved can be detrimental to students' conceptions of natural selection" (Mulvey et al., 2016, S. 516).

Aus einer kognitiven Perspektive ist das nachvollziehbar: Zuerst muss Wissen entwickelt werden, bevor man dann über dieses reflektieren kann. Mulvey et al. (2016) führen aber noch einen motivationalen Aspekt an, den Grad der Kontextualisierung nur schrittweise zu erhöhen.

> "Initial learning in lessons embedded in rich disciplinary contexts may discourage teachers from attempting inquiry and/or NOS instruction; highly contextualized NOS instruction can be difficult for teachers to implement. Thus, NOS instruction along a context continuum may strike a balance between accessibility and disciplinary perspective" (Mulvey et al., 2016, S. 516).

Hier wird ein kritischer Punkt im Feld von NOS deutlich. Es wurde viel zu den Vermittlungsgängen, den Dimensionen von NOS und zur Kontextualisierung der Vermittlung geforscht und publiziert. Insbesondere die neueren Studien arbeiten jedoch mit Lehrkräften. Es fehlen demnach Forschungsarbeiten zur Implementierung in Klassen und somit zur Elaborierung von NOS-Konzepten bei Schülerinnen und Schülern. Hierzu gehören Untersuchungen zur Länge und der Qualität der NOS-Vermittlung sowie das Ausmaßder Kontextualisierung aber auch die Verbindung zwischen NOS- und fachlich orientierten Stunden. Dieses Problem fassen Mulvey et al. (2016) pointiert zusammen:

> "Finally, while understanding how teachers learn about and teach NOS, the ultimate question is how does such instruction impact student learning?" (Mulvey et al., 2016, S. 517).

Folglich besteht die Frage, ob eine Intervention, die den *explizit-reflektierenden* Ansatz

kombiniert mit einer experimentellen Problemstellung einsetzt, in der Lage ist, NOS-Konzepte von Schülerinnen und Schülern kontextualisiert zu elaborieren.

2.4.3 Die adäquate Erfassung von NOS-Konzepten

In den letzten 60 Jahren wurden mehr als 20 standardisierte Testinstrumente entwickelt, um NOS-Konzepte der Schülerinnen und Schüler zu erfassen (Harrison, Duncan Seraphin, Philippoff, Vallin & Brandon, 2015). Das Testformat besteht jeweils aus Auswahlelementen in Form von Multiple-Choice-Antworten oder in der Umsetzung von Likertformaten. Im Feld von NOS wurden jedoch immer wieder Einwände gegen standardisierte Testinstrumente erhoben, was zu einer Dominanz von standardisierten Interviewleitfäden – vor allem des Views about Nature of Science (VNOS) - geführt hat (Lederman et al., 2002). Die Kritik an der Verwendung von standardisierten Instrumenten bezieht sich dabei auf die Annahme, dass über diese aussagekräftige Schlussfolgerungen in Bezug auf die Art und der Weise der NOS-Konzepte der Lernenden ermöglicht werden.

Urhahne et al. (2008) stellen jedoch zusammenfassend fest, dass die Kritik an der standardisierten Erhebung von NOS-Konzepten nicht mehr auf neuere Testinstrumente zutrifft. Für Urhahne et al. (2008) steht aber auch fest, dass die Kritik "[. . .] selbst bei äußerst solider Testkonstruktion nicht vollständig zurückgewiesen werden kann" (80). Diese letzte Ungewissheit beruht vor allem auf dem tatsächlichen Verständnis, was sich durch Items bei Schülerinnen und Schülern einstellt.

> "Researchers also found that the traditional instruments fail to detect either the subjects' perceptions and interpretations of the test items or their underlying reasons for making choices" (Chen, 2006, S. 804).

Unterstützt wird diese Annahme durch eine Studie von Aikenhead (1979) im Feld von NOS und im Feld des konzeptionellen biologischen Fachwissens von Unger (2018).

Warum werden dennoch quantitative Instrumente im Rahmen dieser Studie verwendet? Im Wesentlichen lässt sich diese Frage auf zwei Arten beantworten. Harrison et al. (2015) weisen darauf hin, dass die NOS-Dimensionen nicht theoretisch, sondern auch empirisch miteinander verbunden sind. Selbst eine größtmögliche Sorgfalt bei der Formulierung der Items und der Zusammenführung dieser in Skalen, kann nicht verhindern, dass diese sich tatsächlich distinkt voneinander verhalten. Qualitative Instrumente können die Frage nach dem Zusammenhang der NOS-Inhaltsbereiche jedoch gar nicht beantworten, weil einerseits die Fallzahlen zu gering sind und andererseits das Datenformat aus Interviews keine Analysen ohne die Bildung von Skalen aus diesen Daten zulässt.

Ein zweiter, und für diese Arbeit wesentlicher Grund für den Einsatz von standardisierten Testinstrumenten, besteht in der Frage nach einer möglichen Abhängigkeit von Daten aus NOS-Instrumente und den Lernendenmerkmalen. Dieser Zusammenhang lässt sich ebenfalls nicht interpretativ lösen, weil hierfür wiederum Zusammenhangsanalysen zwischen den Interviewdaten und den Lernendenmerkmalen vorgenommen werden müssten. In der Folge besteht erneut die Frage nach einer ausreichend hohen Fallzahl. Zusammenfassend kann daher mit Urhahne et al. (2008) argumentiert werden, dass der Einsatz von Likertformaten mit einer Zustimmungsskala Items weniger anfällig gegen eine Fehlinterpretation der Testprobanden macht, als es bei Multiple-Choice-Formaten der Fall ist. Daher muss

das Risiko der Fehlinterpretation und der Nutzen durch den Einsatz eines standardisierten Instruments in einem akzeptablen Verhältnis stehen.

2.5 Zusammenfassung des theoretischen Rahmens

Ein produktiver Umgang mit Lernendenmerkmalen zeichnet sich in erster Instanz dadurch aus, dass diese als Ressource für den Unterricht begriffen werden. Die Studienlage zeigt, dass die Einstellungen von Lehrenden hierbei eine zentrale Rollen für die *Umsetzung eines inklusiven Unterrichts* im Sinne einer *Inclusive Education* darstellen. Auf der Ebene der Unterrichtsgestaltung werden Differenzierungsmaßnahmen als wirksam beschrieben (z.B. Eselsbrücken: (Hattie, 2014; Scruggs, Mastropieri, Berkeley & Graetz, 2010; Steele, 2008; Therrien et al., 2011; Villanueva & Hand, 2011); Aufgabenbearbeitungsstrategien: (Browder et al., 2012); Peer-tutoring: (Jones & Sterling, 2011; Scruggs & Mastropieri, 2007)). Differenzierungsmaßnahmen ermöglichen es, Aufgaben zu gestalten, die hinsichtlich des Schwierigkeitsgrades differenziert sind und gleichzeitig eine Arbeit am gemeinsamen Lerngegenstand für alle Lernenden ermöglichen (Abels & Markic, 2013; Tobin & Tippett, 2013). Gerade letzter Punkt ist in der Diskussion um Differenzierungsmaßnahmen entscheidend. Werden lediglich die Oberflächenmerkmale des Unterrichts differenziert, ergeben sich nur geringe Effekte (Hattie, 2014).

Eine Möglichkeit zur Differenzierung von Unterricht besteht im UDL (CAST, 2018). UDL bietet diverse *Zugangsmöglichkeiten* zu einem Unterrichtsgegenstand und formuliert ein zentrales Unterrichtsziel. Die Zugangsmöglichkeiten ergeben sich sowohl auf der Ebene des Inhalts durch verschiedene Repräsentationsformen und assistive Systeme als auch auf der motivationalen Ebene, indem die Bedeutung des Lerninhalts durch *Alltagsnähe* und *übertragbarkeit* hergestellt sowie *Autonomieerleben* durch Wahlmöglichkeiten ermöglicht wird. Schließlich bieten prozedurale Unterstützungsmöglichkeiten das Lernen im eigenen Tempo und die Möglichkeit zur *Selbstkontrolle*. UDA hingegen formuliert klare Ansprüche an das Assessment und denkt – wie UDL den Lernprozess – den Leistungsprozess vor dem Hintergrund der Zugänglichkeit. Ziel ist es, die KIV durch Testadaptionen zur reduzieren.

UDL bietet Anknüpfungspunkte für experimentelles Arbeiten, um naturwissenschaftliche Inhalte in diversen Lerngruppen zu vermitteln (Therrien et al., 2011, 2014; Villanueva et al., 2012). Hierbei werden die Lernenden aufgefordert, eine Problemstellung durch die Ausarbeitung und Durchführung von Experimenten zu beantworten. Durch diese Form des Unterrichts rückt das Experiment in den Vordergrund und kann, je nach Schwerpunktsetzung, fachliche Inhalte oder auch Reflexionsprozesse fokussieren. Das selbstgesteuerte Lernen knüpft damit sowohl an die motivationalen, als auch an die prozeduralen Aspekte von UDL an. über geeignete Reflexionspunkte (z.B. in Form von Self-Assessments) können die von der Lehrkraft intendierten Ziele durch die Lernenden explizit reflektiert werden. Schließlich kann über das Prinzip der multiplen Repräsentationsformen auch die Teilchenebene zugänglicher gestaltet werden, um Lernprozesse zu naturwissenschaftlichen Theorien zu ermöglichen. Werden darüber hinaus alltagsnahe Problem- und Fragestellungen experimentell und theoretisch zugänglich gestaltet, fördert dies die Motivation, weil das fachliches Wissen Relevanz in der Alltagswelt der Lernenden erfährt.

Für diese Studie wurde die Frage *Ist gleich viel auch gleich schwer?* im Sinne von UDL zu mehreren experimentellen Lernumgebungen aufgearbeitet. Die explizite Reflexion, die

notwendig ist, um NOS-Konzepte zu elaborieren, ermöglicht die Abstraktion vom fachlichen Lerninhalt zum Zusammenhang von Volumen und Masse, der auch genutzt werden könnte, um das Dichtekonzept einzuführen. So angelegt, fokussiert die Lernumgebung auf den *Zweck des Experiments* (Testen von Hypothesen) und die *Experimentierplanung* (Theoretische Konzeption von Experimenten). Die alltagsweltliche Kontextualisierung im Rahmen eines Experiments kommt der Forderung nach, dass NOS-Konzepte nicht isoliert elaboriert werden sollten.

Die Betonung dieser Elemente begegnet dem Umstand, dass gerade jüngere Schülerinnen und Schüler Vorstellungen zeigen, die das Experimentieren als einen planlosen Prozess beschreiben. Insgesamt fokussiert diese Studie daher auf NOS-Konzepte zur *Rechtfertigung von naturwissenschaftlichem Wissen*. Hierzu sind vor allem Argumentationsfähigkeiten erforderlich, die mit den Aspekten der naturwissenschaftlichen Grundbildung korrespondieren.

Insgesamt stellen UDA und UDL vielversprechende Ansätze dar. Wird insbesondere UDL mit Ansätzen zum *Forschenden Lernen* verknüpft, können im naturwissenschaftlichen Unterricht alltagsnahe Fragestellungen experimentell und theoretisch zugänglich gestaltet werden. Dies wird möglich, weil UDL die explizite Reflexion im Vermittlungsgang vorsieht. In dieser Weise werden auch die Forderungen nach einem guten, inklusiven Unterricht von Fischer et al. (2014), Klieme & Rakoczy (2008) und Sliwka (2012) erfüllt (Tab. 2.1).

3 Fragen und Ziele der Studien

3.1 Ziele der Studien

Aus dem theoretischen Rahmen des Projekts leiten sich verschiedene Zielstellungen ab. Für die konzeptionelle Gestaltung der Lernumgebung sollen alltagsnahe Fragen mit Experimente untersucht werden. Das Universal Design for Learning (UDL) dient hierzu als theoretischer Rahmen zur Planung und Gestaltung der Lernumgebungen. Im Fokus steht die explizit-reflektierende Vermittlung zu Elaborierung von Nature of Science (NOS) Konzepten mit besonderen Fokus auf die *Rechtfertigung des naturwissenschaftlichen Wissens* (Abschnitt 2.4). Dies begründet sich vor allem in der Bedeutung der datenbasierten Argumentation sowie dem Zusammenhang zwischen dem Design eines Experiments und der sich hieraus ergebenden Art und Weise der Datengewinnung.

In einem ersten Schritt müssen verschiedene Inhaltsrepräsentationsformen für den Lerninhalt entwickelt und die Gleichwertigkeit dieser überprüft werden. In einem zweiten Schritt werden auf Basis dieser Daten Lernumgebungen entwickelt, die einerseits alle UDL-Prinzipien umfassen (UDL-Lernumgebung), um andererseits den Vergleich zu einer Lernumgebung ziehen zu können, die auf das *Prinzip der multiplen Repräsentation* (MR-Lernumgebung) fokussiert (Abschnitt 2.3.1). Außerdem ergibt sich auch aus dem Forschungsfeld zu UDL das Problem der fehlenden Evidenz zu primären Bildungseffekten (Al-Azawei et al., 2016; Capp, 2017; Rao et al., 2014; Serenelli & Mangiatordi, 2013). Daher wird drittens das Universal Design for Assessment (UDA) (Beddow, 2011; Lovett & Lewandowski, 2015; Thompson et al., 2002) zur Erfassung von NOS-Konzepten verwendet, um möglichst adäquat eventuelle primäre Bildungseffekte beobachten und beschreiben zu können. Dahinter steht viertens die Frage nach der *Testzugänglichkeit* und der *konstruktirrelevanten Varianz* (KIV) (Abschnitt 2.3.2). Hierzu wird auch der *Einfluss von Lernendenmerkmalen* (die Lesefähigkeit, der sozioökonomische Status, die Intelligenz, der sonderpädagogische Unterstützungsbedarf, die kognitive Aktivierung und der wahrgenommene Nutzen der Lernumgebungen) auf die Elaborierung von NOS Konzepten bestimmt (Conley et al., 2004; Greene, Cartiff & Duke, 2018; Kampa et al., 2016).

Für die Analysen muss in einem ersten Schritt überprüft werden, ob die gleichen latenten Konstrukte in einem adaptierten und einem nicht-adaptierten Assessment für NOS-Konzepte abgebildet werden (2. Precisely defined constructs; Abschnitt 2.3.2) (Lovett & Lewandowski, 2015). Daran anschließend soll zweitens überprüft werden, ob die *Testzugänglichkeit* durch die Operationalisierung von UDA ermöglicht und die KIV reduziert wurde (3. Accessible, non-biased items) (Beddow, 2011). Drittens sollen die Assessments mit Bezug auf die Lernumgebung auf ihre *instruktionale Sensitivität* hin überprüft werden (Polikoff, 2010). Viertens werden statistische Methoden unter Einbezug der Lernen-

denmerkmale eingesetzt, die sowohl die interindividuelle, als auch die intraindividuelle Entwicklung aufklären. Hierzu zählt auch die Frage nach dem *Differential Boost*, der durch die Adaptionen möglicherweise vorliegt (Elliott et al., 2018; Phillips, 1994). Weitergehend sollen die Effekte mit Bezug auf oder durch die Hintergrundvariablen analysiert werden, um einerseits Aussagen zum *Lernen für alle* als auch für das Individuum treffen zu können.

Zusammenfassend zielt die Studie auf die Adressierung von Lernendenmerkmalen über universell designte Lernumgebungen und Assessments. Für die Umsetzung und Evaluation sind folgende Aspekte relevant.

1. Der Vergleich zweier potenziell barrierefreier Lernumgebungen (UDL-Lernumgebung vs. MR-Lernumgebung)
2. Der Vergleich der Messung der NOS-Konzepte über zwei verschiedene Testformate (Testzugänglichkeit und KIV)
3. Die Untersuchung des Einflusses der Hintergrundvariablen auf die Messung und die Elaborierung von NOS-Konzepten (*Differential Boost* und *Lernen und Lehren für alle*)

3.2 Fragestellungen der Studien

Aus den Zielstellungen ergibt sich die für das Projekt leitende Forschungsfrage:

> Inwiefern lassen sich Effekte zur Elaborierung von NOS-Konzepten beim Einsatz von universell designten Lernumgebungen und Assessments unter Berücksichtigung von Lernendenmerkmalen beobachten?

In einer qualitativen Vorstudie wurden drei Lernumgebungen mit je einer Inhaltsrepräsentationsform (Pop-Up-Text, Comic und Video) entwickelt und im Prä-Post-Design qualitativ evaluiert. Die drei folgenden Fragen sind hierfür leitend:

1. Inwiefern führen die Lernumgebungen mit je einer Repräsentationsform zu unterschiedlichen Elaborierungen von NOS-Konzepten?
2. Inwiefern führen die entwickelten Lernumgebungen zu einer Elaborierung von NOS-Konzepten bei allen Schülerinnen und Schülern?
3. Inwiefern lassen sich Unterschiede in der Elaborierung von NOS-Konzepten im Vergleich zwischen den Schülerinnen und Schülern mit und ohne sonderpädagogischen Unterstützungsbedarf beobachten?

Auf Basis der Vorstudie wurde eine ganzheitliche UDL-Lernumgebung gestaltet, die die drei Inhaltsrepräsentationsformen umfasst sowie eine video-basierte MR-Lernumgebung. Im Anschluss hieran wurden in einer quantitativen Hauptstudie die Ergebnisse der Vorstudie überprüft.

In einem ersten Schritt werden die Assessments auf ihre Reliabilität, bezüglich der Messinvarianz und der instruktionalen Sensitivität hin überprüft.

> Inwiefern unterscheiden sich das Original- und das UDA-Assessment in der Erfassung von NOS-Konzepten ...
>
> 4. ... hinsichtlich der internen Konsistenz und der Skalenstruktur?

5. ... hinsichtlich der longitudinalen Messinvarianz und der instruktionalen Sensitivität?
6. ... hinsichtlich der Gruppenabhängigkeit (Lernendenmerkmale) von Items (Testzugänglichkeit)?
7. ... hinsichtlich eines *Differential Boost* durch die Testadaptionen?

In einem zweiten Schritt werden die instruktionalsensitiven Skalen verwendet, um die Effekte der UDL- und der MR-Lernumgebung zu untersuchen. Dies geschieht ohne und unter Kontrolle der Lernendenmerkmale. Über die Analysen der Kovarianzen in den Strukturgleichungsmodellierungen, Effektstärken basierend auf Mittelwertsverschiebungen und über Subgruppenanalysen zur internen Konsistenz der Skalen bei Risikolernenden und den übrigen Schülerinnen und Schülern wird einem möglichen *Differential Boost* nachgegangen.

Inwiefern unterscheiden sich die UDL-Lernumgebung und die MR-Lernumgebung...

8. ... in der interindividuellen Elaborierung von NOS-Konzepten unter Einbezug von Lernendenmerkmalen?
9. ... in der intraindividuellen Elaborierung von NOS-Konzepten unter Einbezug von Lernendenmerkmalen?

4 Methodisches Vorgehen

4.1 Beschreibung der Lernumgebungen

In einem ersten Schritt wurden zur Konzeption der Lernumgebungen drei verschiedene Inhaltsrepräsentationsform ausgewählt. Die erste Inhaltsrepräsentationsform stellt ein Video dar. Hierin bestand die Idee, dass dies vermutlich die dominante Inhaltsrepräsentationsform von Inhalten für die Altersstufe der Probanden sein wird. Eine weitere Inhaltsrepräsentationsform stellt ein Comic dar. Hierbei bestand die Vermutung, dass die Text-Bild-Integration als Geschichte ähnlich motivierend sein kann wie ein Video. Schließlich wurde eine interaktiver Pop-Up-Text erstellt, der über Schaltflächen bei Bedarf zusätzliche Informationen im Sinne einer Lernleiter bereitstellt. Diese Inhaltsrepräsentationsform wurde ausgewählt, weil sie ähnlichkeiten zum Schulbuchtext aufweist (Abb. 4.1).

In der Vorstudie bestand die Frage, ob verschiedene Inhaltsrepräsentationsformen (Pop-Up-Text, Video und Comic) unterschiedlich in der Elaborierung von Nature of Science (NOS) Konzepten wirken. Nach der Vorstudie wurden theorie- und evidenzbasiert diese Inhaltsformen in einer Universal Design for Learning (UDL)-Lernumgebung zusammengeführt und durch weitere Strukturierung, wie Verlinkungen zwischen den Teilen des E-Books (Aufgabenstellung, Experiment und Self-Assessment) erweitert (Abb. 4.2 und Abschnitt 7.1).

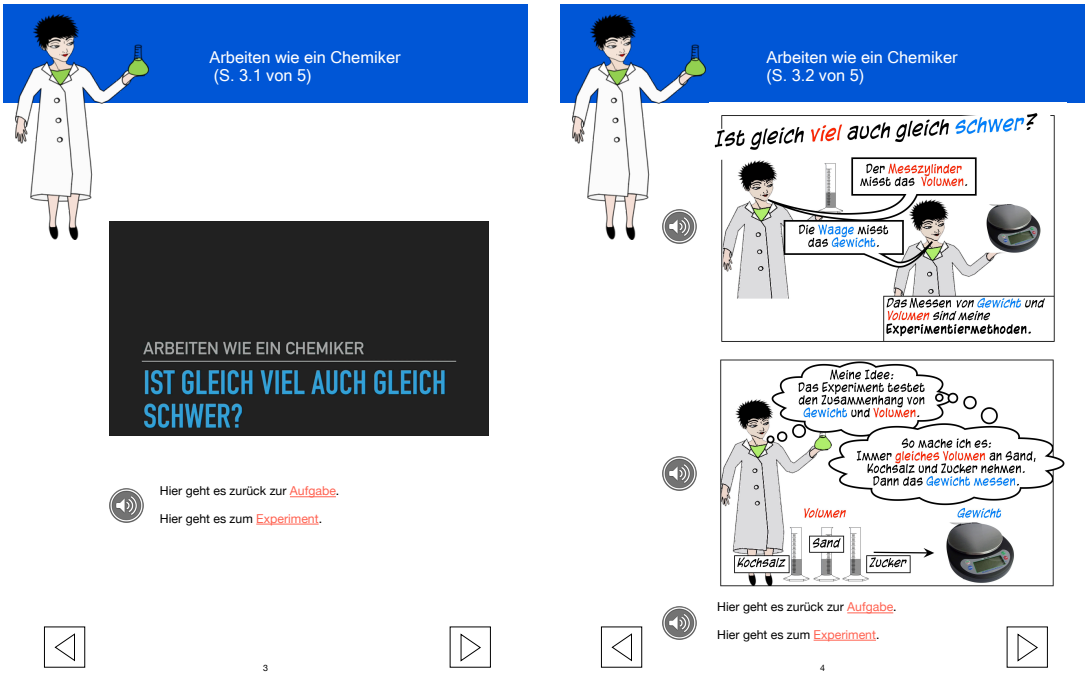


Abbildung 4.1: Inhaltsrepräsentationsformen des Lerninhalts der Lernumgebungen.

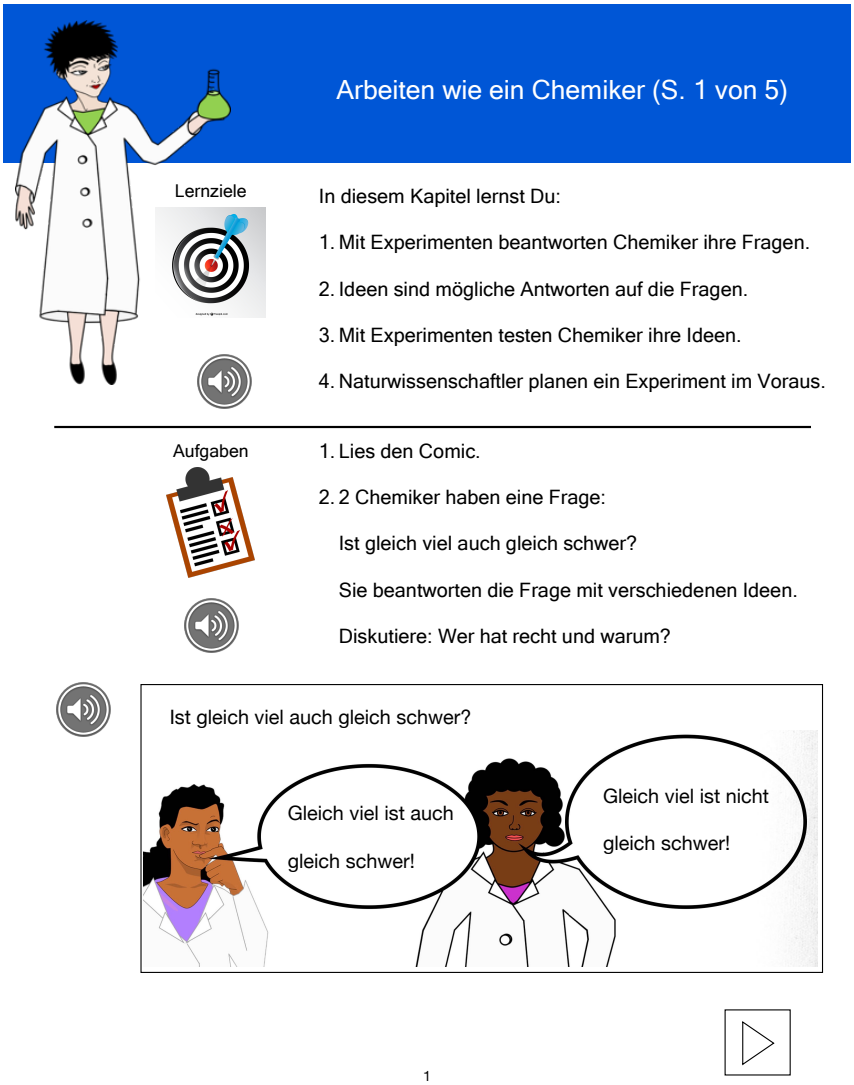


Abbildung 4.2: Erste Seite der UDL-Lernumgebung.

Die zweite Lernumgebung (MR-Lernumgebung) umfasst das gleiche Video, wie die UDL-basierte Lernumgebung und ist darüber lediglich mit der gleichen Aufgabenstellung versehen (Abb. 4.3 und Abschnitt 7.2). Sie verfügt darüber hinaus über keine weitere Funktionen oder Adaptionen und ist damit als Lernumgebung mit dem Schwerpunkt zum *Prinzip der Multiplen Repräsentation* ausgelegt (King-Sears et al., 2015). Das Video lässt sich beliebig oft abspielen.

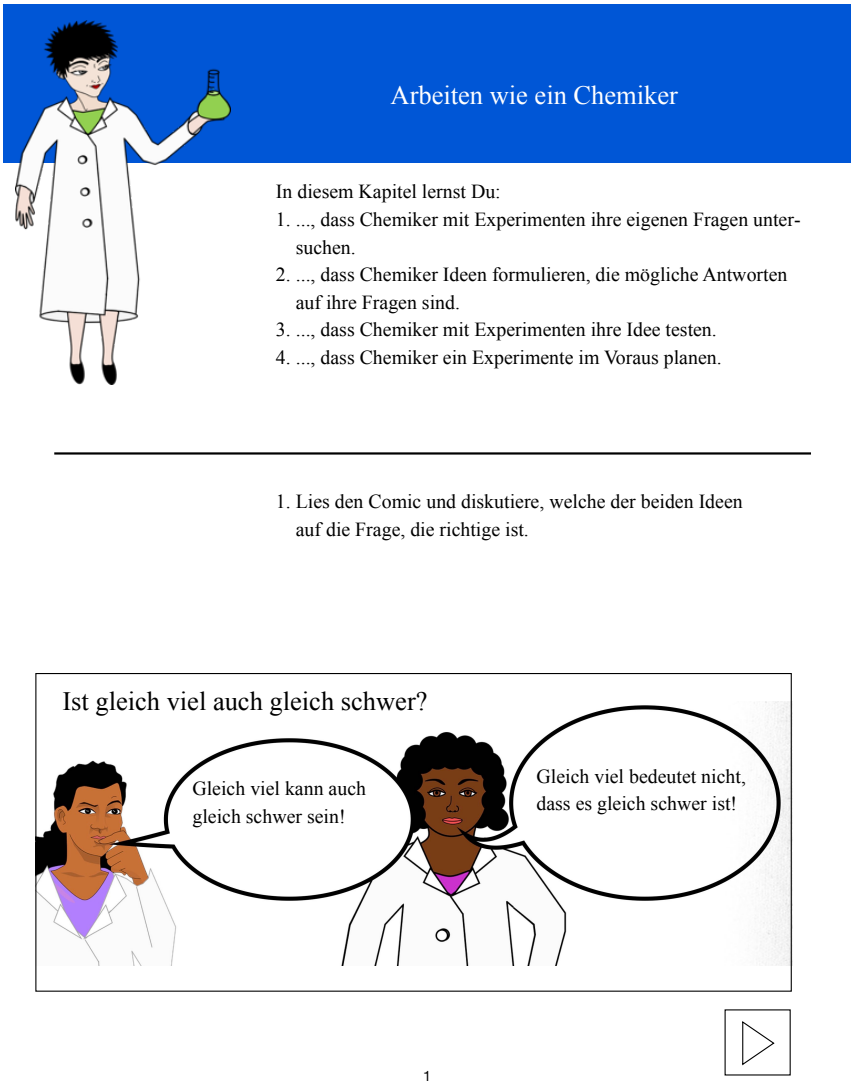


Abbildung 4.3: Erste Seite der MR-Lernumgebung.

Für die Ausgestaltung der Lernumgebung wurden Elemente aus den Forschungsfeldern von UDL (Abb. 2.3.1) und NOS (Abb. 2.4) angewendet. Hierzu gehörte die Arbeit in Kleingruppen (Henderson, MacPherson, Osborne & Wild, 2015; McGinnis, 2013; Scruggs et al., 2010); die Schülerinnen und Schüler bearbeiteten die Lernumgebungen in der Regel in Vierergruppen. Die Lernumgebungen sahen theoretisch gerahmte Hands-on Aktivitäten vor (Hodson, 2014); die Schülerinnen und Schüler planten und führten ein Experiment durch und wurden dabei durch zwei konkurrierende Hypothesen geleitet. Hierüber führten die Lernumgebungen in ein explizit-reflektierendes Arbeiten ein (Bell, Mulvey & Maeng, 2016). Die Schülerinnen und Schüler wurden durch die Inhaltsrepräsentationsform aufgefordert Variablen explizit zu ändern bzw. konstant zu halten. Dies geschieht immer mit Blick auf die untersuchte Hypothese. Am Ende sollte begründet werden, welche Hypothese zugestimmtn werden kann. In der UDL-Lernumgebung geschah dies zusätzlich über das Self-Assessment.

Die Operationalisierung der UDL-Richtlinien stützt sich auf Forschungsergebnisse aus der Testentwicklung und -evaluation (Salvia, Ysseldyke & Witmer, 2017; Thompson et al., 2002) sowie aus der Forschung zu digitalen Lernwelten (Clark & Mayer, 2016) (Tab. 4.1).

Tabelle 4.1: Operationalisierung der UDL-Lernumgebung.

Operationalisierung	UDL-Richtlinie
MS Sans Serif 18	1.
Zeilenabstand 2,0	1.
Leichte Sprache	1./2.
Piktorale Unterstützung zur Unterscheidung von Textarten (Lernziele, Aufgabenstellungen, Lerninformationen)	2.
Auswahl der Inhaltsrepräsentationsform (Pop-Up-Text, Comic, Video)	3./7.
Vorlesefunktion	3.
Seitenorganisation	8./9.
Arbeit mit einer Checkliste	6.
Self-Assessment zu den Lerninhalten	5./9.
Arbeit an Realobjekten	4.
iPad-basiert	4.
Gruppenarbeit/Peer-Tutoring	5.

UDL-Richtlinie:
[1] Perzeption des Lerninhalts
[2] Darstellungen von sprachlichen und symbolischen Informationen des Lerninhalts
[3] Verständnis des Lerninhalts
[4] Interaktion mit dem Lerninhalt
[5] Arbeit mit dem und Kommunikation über den Lerninhalt
[6] Verarbeitung des Lerninhalts
[7] Angebote zur Weckung des Lerninteresses
[8] Erhaltung eines engagierten Lernens
[9] Selbstreguliertes Lernen

Die Erstellung der Lernumgebung erfolgte über iBooks-Author im E-Bookformat (Apple Inc., 2017). Es sei dabei angemerkt, dass sich eine konkrete Adaption bzw. Operationalisierung durchaus mehreren UDL-Elementen zuordnen lässt und deshalb nicht immer bis ins Letzte eindeutig ist.

Alle Lernumgebungen fokussierten inhaltlich auf die Frage *Ist gleich viel auch gleich schwer?*, die über die Planung und Durchführung eines Experiments bearbeitet wird. Im Sinne von UDL wurde ausdrücklich diese alltagsnahe Frage nach dem Zusammenhang von Volumen und Masse gewählt. Für die sprachliche Ausgestaltung wurde ein theoretischer Rahmen zur Leichten Sprache verwendet (Inclusion Europe, 2016). Die Regelliste umfasst z.B. die Wahl von kürzeren Sätzen oder Wörtern oder die Verwendung der aktiven Grammatikform. Die Wortwahl sollte möglichst alltagsnah sein. So wurden die Worte *Volumen* und *Masse* beispielsweise durch *Menge* und *Gewicht* ersetzt. Dies begründet sich in der Anlage der Lernumgebungen. Beide verfolgen die Förderung von NOS- und nicht die von fachlichen Konzepten. Hierzu wurde über die Datenbank von Duden.de die Alltagsnähe überprüft (Bibliographisches Institut, 2018). Das Wort, z.B. *Masse*, wird in die Datenbanksuche eingetragen. Anschließend wird eine Beschreibung des Bedeutungsumfelds angezeigt und eine Häufigkeit des Worts in der Alltagssprache in fünf Stufen ausgegeben (Stufe fünf "bedeutet, dass das Wort zu den 100 häufigsten Wörtern im Dudenkorpus gehört"; "Stufe

eins bedeutet, dass das Wort jenseits der Top 100 000 liegt und nur selten oder gar nicht im Dudenkorpus belegt ist" (Bibliographisches Institut, 2018)). Da *Masse* sowohl eine physikalische Größe, als auch eine Menge darstellen kann und darüber zusätzlich einen stofflichen Aspekt aufweist, wurde es für die Lernumgebung nicht verwendet.

Um die Aufmerksamkeit auf die NOS-Ebene zu richten, wurden die Experimentiermaterialien so ausgewählt, dass die Experimente anhand der beiden Hypothesen ("Gleich viel ist nicht unbedingt gleich schwer" und "Gleich viel ist auch gleich schwer") planbar sind. Sand, Salz und Zucker dienten neben Messzylindern aus Plastik, Waagen und Spatellöffeln der Planung des Experiments. Um Identifikationsmöglichkeiten zu schaffen, beginnt die Lernumgebung mit zwei Chemikerinnen, die über die beiden Hypothesen streiten und für je eine Hypothese bildlich stehen. Es folgt eine Einführung in die Experimentiermaterialien. Nach dem Experiment sollte datenbasiert entschieden werden, welche der beiden Hypothesen zutrifft und welche falsifiziert wurde. Die Aufgabenstellung zielt auf den *Zweck des Experiments* und die *Experimentierplanung*, welche mit der NOS-Dimension der *Rechtfertigung* korrespondieren.

4.2 Design der Studien

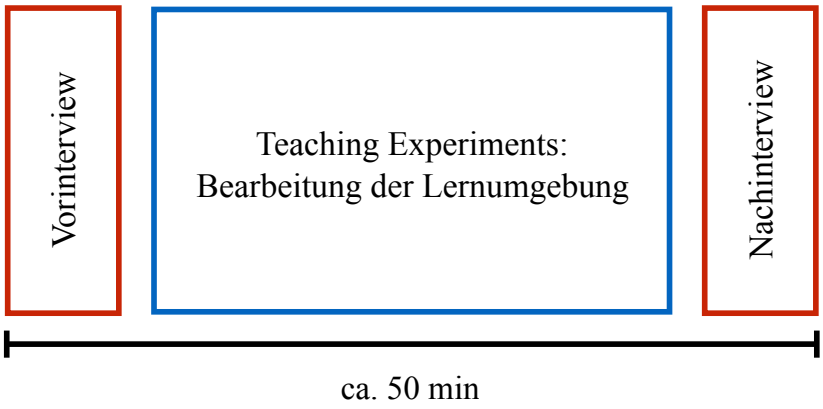


Abbildung 4.4: Das Design der Vorstudie.

Für die ***qualitative Studie*** wurde ein Design mit einem Vor- und einem Nachinterview gewählt. Ziel der Studie war es, die Vergleichbarkeit der Inhaltsformen zu untersuchen. Die Studienteilnehmer wurden hierzu jeweils einzeln interviewt. Die Schülerinnen und Schüler führten anschließend in Gruppen à vier Personen die Bearbeitung der Lernumgebung durch (Abb. 4.4). Jede der drei Lernumgebungen wurde gleich oft eingesetzt. Die Arbeit mit der Lernumgebung umfasst ca. eine halbe Stunde. Die Vor- und Nachinterviews dauerten ca. zehn Minuten je Zeitpunkt.

Für die ***quantitative Studie*** wurde ein 2x2-Between-Subject-Design gewählt. Das Design zielt einerseits auf die Unterschiede in den Lernumgebungen und den Assessments. Darüber hinaus werden die Ergebnisse über Lernendenmerkmale kontrolliert. Dies bietet den Vorteil, dass Gruppen unter Hinzunahme diverser Daten verglichen werden können. Ein Nachteil an diesem Design besteht in der notwendigerweise großen Stichprobe, um jede Gruppe ausreichend stark zu gestalten.

Tabelle 4.2: Das Design der Hauptstudie.

Assessment	Lernumgebung	
	UDL-LU	MR-LU
UDA-Assessment	Gruppe 1	Gruppe 2
Originalassessment	Gruppe 3	Gruppe 4

Anmerkung:
LU: Lernumgebung
MR: multiple Repräsentation
UDL: Universal Design for Learning
UDA: Universal Design for Assessment

Entsprechend der Forschungsfragen bilden die Lernumgebungen sowie die Assessmentformen je eine Untersuchungsbedingung in zwei Kategorien ab (Tab. 4.2). Die Lernenden wurden in Gruppen à vier Personen einer der vier Untersuchungsbedingungen zufällig zugewiesen. So konnte sichergestellt werden, dass jede Untersuchungsbedingung in jeder teilnehmenden Klasse mindestens einmal vertreten war. Die gesamte Intervention umfasste ca. 90 Minuten.

4.3 Datengewinnung und Auswertung der Vorstudie

Zur Beantwortung der Forschungsfragen 1-3 (Abschnitt 3.2) wurden vor und nach den Vermittlungsversuchen (Komorek & Duit, 2004) leitfadengestützte Interviews durchgeführt (Schmidt, 2013). Dieses Vorgehen bietet den Vorteil einer Standardisierung der Interviews bei gleichzeitiger Möglichkeit spontane Nachfragen zu stellen. Außerdem sind die Vermittlungsversuche sowohl didaktisch als auch diagnostisch gerahmt, weil der Leiter der Untersuchung als Lehrkraft und als Interviewer auftritt. Der Leitfaden umfasste Fragen zur Konzeptentwicklung, zur Lernumgebung selbst sowie zu ihrer Nutzbarkeit für das eigene Lernen (Abschnitt 7.5). Weiterhin wurden einige Lernendenmerkmale bezüglich des Geschlechts, dem Jahrgang und einem eventuell vorhandenen Förderbedarf aufgenommen.

Die Auswertung der Daten aus der Vorstudie erfolgt über die Qualitative Inhaltsanalyse (QI) (Mayring, 2013). In der Variante der inhaltlich strukturierenden QI wird anhand von Kategorien das Material analysiert (Kuckartz, 2012). In einem ersten Schritt werden Hauptkategorien kodiert, die theoretisch abgeleitet sind. Die Anzahl der Kategorien in dieser ersten Phase ist meist relativ gering. In einem zweiten Schritt werden dann die Kategorien am Material weiterentwickelt und ausdifferenziert. Allerdings lässt sich die Anzahl der Kategorien auch am Material induktiv erweitern, was insbesondere der Suche nach möglichen unbekannten Schülervorstellungen nachkommt.

Im Rahmen dieses Projekts wurden die Kategorien mit Hilfe von Carey et al. (1989) gebildet (Abschnitt 7.4). In Kombination beider theoretischer Rahmen können äußerungen zum *Zweck des Experiments* und zur *Planung um das Experiment* zur Beantwortung der Forschungsfragen gewonnen werden. Beide Kategorien stehen in Beziehung zur NOS-Dimension *Rechtfertigung des naturwissenschaftlichen Wissens*, indem sie die Bedeutung von Beobachtungen und Experimenten über die Konstruktion von naturwissenschaftlichen

Untersuchungen, wie dem Experiment, thematisieren. Schlussendlich umfassen sie auch die evidenz- und theoriebasierte Argumentation zur Auswertung der gewonnenen Daten.

Um eine Interraterreliabilität zu gewährleisten, wurde zehn Prozent des Materials doppelt kodiert. Vor dem Hintergrund zur Erstellung einer inklusiven Lernumgebung, besteht auch ein Interesse daran, die Abhängigkeit der Aussagenkodierung vor dem Hintergrund von weiteren Daten zu untersuchen. Weil jedoch Interviewaussagen und die Lernendenmerkmale weder eine Skalierung noch einen monotonen oder einen konkret-funktionalen Zusammenhang aufweisen, werden diese als nominale Merkmale bezeichnet und aufgefasst (Burkschat, Cramer & Kamps, 2012). Für die Beschreibung eines Zusammenhangs sind daher die absoluten und relativen Häufigkeiten der Kodierungen in Bezug auf die weiteren Daten von Interesse.

über χ^2–Unabhängigkeitstest, kann dann der Zusammenhang von Merkmalen statistisch bestimmt werden. Während der χ^2-Unabhängigkeitstest Auskunft darüber gibt, ob ein statistischer Zusammenhang zwischen zwei Merkmalen vorliegt, helfen Assoziationsmaße, wie Carmérs ϕ, die Stärke des Zusammenhangs zu deuten.

> "Diese Kenngröße der deskriptiven Statistik gibt an, wie unabhängig bzw. abhängig die in der Kontingenztafel erfassten Merkmale sind. $\phi = 0$ entspricht der totalen Unabhängigkeit, während $\phi = 1$ eine größtmögliche Abhängigkeit bedeutet" (Falk, Hain, Marohn, Fischer & Michel, 2014, S. 194).

4.4 Beschreibung der Instrumente der Hauptstudie

4.4.1 Konzeption der NOS-Assessments

Das Originalassessment umfasst vier NOS-Dimensionen (Kampa et al., 2016). Hierzu zählten die *Sicherheit des naturwissenschaftlichen Wissens* (sieben Items), was zwar relativ verlässlich und dauerhaft sein kann, aber stets Vorläufigkeitscharakter aufweist. Demnach können auch verschiedene Theorien ein und dasselbe Phänomen erklären. Die *Entwicklung von naturwissenschaftlichem Wissen* (acht Items) stellt ebenfalls eine Skala des Assessments dar. Demzufolge unterliegt das naturwissenschaftliche Wissen einer fortwährenden Entwicklung und ist Veränderungen ausgesetzt. Die *Rechtfertigung des naturwissenschaftlichen Wissens* (sieben Items) bezieht sich vor allem auf die evidenz- und theoriebasierte Argumentation und stellt eine dritte NOS-Skala dar. Sie umfasst die Bedeutung von Beobachtungen und Experimenten sowie die Begründungen zur Generierung von naturwissenschaftlichem Wissen. Die letzte Skala bezieht sich auf die *Herkunft des naturwissenschaftlichen Wissens* (fünf Items). So kann naturwissenschaftliches Wissen nicht ausschließlich von Naturwissenschaftlern gebildet werden, sondern auch von Schülerinnen und Schülern. Hierzur sind lediglich Neugierde und konkrete Fragen notwendig.

Der erste Schritt zu einem nach dem UDA adaptierten NOS-Instrument lag in der Vereinfachung der Sprache des Originals mit den beschriebenen Skalen. Hier wurden ebenso wie bei den Lernumgebungen die Wörter überprüft (Bibliographisches Institut, 2018). So wurde auch hier das Konzept von Leichter Sprache verwendet (Inclusion Europe, 2016). In einem nächsten Schritt wurde anschließend der vereinfachte Wortlaut in verschiedenen Runden von Experten für Germanistik (zwei), für sonderpädagogische Förderung (zwei) und natur-

wissenschaftlichen Unterricht (zwei) evaluiert. Alle Experten stammen aus kooperierenden Fachbereichen des Instituts für Didaktik der Naturwissenschaften der Leibniz Universität Hannover. Dieses Expertenrating sollte sicherstellen, dass der Kerngedanke des Items trotz Umformulierung erhalten bleibt, sodass beide Assessments inhaltlich vergleichbar sind. Zur besseren Vergleichbarkeit wurden positive oder negative Formulierungen beibehalten, obwohl Hinweise darauf bestehen, dass negative Formulierungen insbesondere für Schülerinnen und Schüler mit geringem sozioökonomischen Hintergrund und geringen Lesefähigkeiten weniger leicht verständlich sind (Salas-Wright, Olate & Vaughn, 2013). Konnte keine adäquate übersetzung gefunden werden, oder entsprach die Originalformulierung bereits den Kriterien, wurde sie beibehalten (Tab. 4.3 und 4.4).

Außerdem wurden für das UDA-basierte Testinstrument größere Textformatierungen verwendet, um die Lesbarkeit zu erhöhen (Thompson et al., 2002). Das Antwortformat wurde in Form von Sternen gestaltet, wie sie aus Onlineshops bekannt sind. Letzteres geschah unter der Annahme, dass dieses Format den Schülerinnen und Schülern aus ihrem Alltag vertraut ist. Die Abb. 4.5 und 4.6 zeigen Bilder der Assessments in der verwendeten App (Wilde, 2018).

Tabelle 4.3: Itemformulierungen der UDA-Assessmentversion.

	Nr.	Formulierung
Herkunft	Item 1	Nur Naturwissenschaftler können die Natur beobachten.
	Item 2	Anfänger können die Natur noch nicht untersuchen.
	Item 3	Nur Naturwissenschaftler können naturwissenschaftliche Theorien entwickeln.
	Item 4	Nur Naturwissenschaftler können naturwissenschaftliche Fragen überlegen.
	Item 5	Anfänger können noch keine naturwissenschaftlichen Fragen stellen.
Sicherheit	Item 1	Das Wissen in den Naturwissenschaften stimmt für immer.
	Item 2	Naturwissenschaftler zweifeln nicht an bewährten naturwissenschaftlichen Theorien.
	Item 3	Ein Naturwissenschaftler findet immer nur eine Lösung für ein Experiment.
	Item 4	Für alle Fragen in den Naturwissenschaften gibt es immer nur eine Lösung.
	Item 5	Naturwissenschaftler können heute nur noch wenig herausfinden, weil fast alles bekannt ist.
	Item 6	Naturwissenschaftler sind sich bezüglich der Naturwissenschaften immer einer Meinung.
	Item 7	Das Beste an den Naturwissenschaften ist, dass viele Probleme nur eine richtige Lösung haben.
Entwicklung	Item 1	Naturwissenschaftler ändern manchmal ihre Meinung darüber, was in ihrem Fach wahr ist.
	Item 2	Naturwissenschaftliche Theorien verändern sich mit der Zeit.
	Item 3	Naturwissenschaftler haben heute einige andere Vorstellungen als früher.
	Item 4	Naturwissenschaftler verändern oder ersetzen naturwissenschaftliche Theorien, wenn neue Beweise vorliegen.
	Item 5	Manchmal ändert sich das Wissen in den Naturwissenschaften.
	Item 6	Auch Forscher haben auf manche Fragen in den Naturwissenschaften keine Antwort.
	Item 7	Neue Entdeckungen können das Wissen von den Forschern ändern.
	Item 8	Das Wissen in den Naturwissenschaften ändert sich manchmal.
Rechtfertigung	Item 1	Naturwissenschaftler führen Experimente mehrmals durch, um das Ergebnis abzusichern.
	Item 2	Wenn Naturwissenschaftler Experimente durchführen, legen Naturwissenschaftler vorher wichtige Dinge fest.
	Item 3	Naturwissenschaftler brauchen klare Ideen, bevor Naturwissenschaftler mit dem Experiment beginnen.
	Item 4	Die Ideen zu naturwissenschaftlichen Experimenten gewinnen Naturwissenschaftler, indem sie neugierig sind und darüber nachdenken, wie etwas funktioniert.
	Item 5	Ein Experiment ist ein guter Weg um herauszufinden, ob etwas wahr ist.
	Item 6	Gute Theorien stützen sich auf Ergebnisse aus vielen verschiedenen Experimenten.
	Item 7	Naturwissenschaftler können auf verschiedenen Wegen ihre Ideen testen.

Anmerkung:
UDA: Universal Design for Assessment

Tabelle 4.4: Itemformulierungen des Originalassessments.

	Nr.	Formulierung
Herkunft	Item 1	Nur Naturwissenschaftler können Naturphänomene beobachten.
	Item 2	Anfänger können noch keine Naturvorgänge beobachten.
	Item 3	Nur Naturwissenschaftler können naturwissenschaftliche Theorien entwickeln.
	Item 4	Nur Naturwissenschaftler können sich naturwissenschaftliche Forschungsfragen überlegen.
	Item 5	Anfänger können sich noch keine naturwissenschaftlichen Forschungsfragen überlegen.
Sicherheit	Item 1	Das Wissen in den Naturwissenschaften ist für alle Zeiten wahr.
	Item 2	Bewährte naturwissenschaftliche Theorien dürfen nicht in Frage gestellt werden.
	Item 3	Es gibt nur die eine Lösung, wenn Naturwissenschaftler einmal das Ergebnis eines Experiments gefunden haben.
	Item 4	Alle Fragen in den Naturwissenschaften haben genau eine Lösung.
	Item 5	In den Naturwissenschaften ist beinahe alles bekannt; es gibt nicht mehr viel, was man herausfinden könnte.
	Item 6	Naturwissenschaftler stimmen immer darin überein, was in ihrem Fach wahr ist.
	Item 7	Das Beste an den Naturwissenschaften ist, dass viele Probleme nur eine richtige Lösung aufweisen.
Entwicklung	Item 1	Manchmal ändern Naturwissenschaftler ihre Meinung darüber, was in ihrem Fach wahr ist.
	Item 2	Naturwissenschaftliche Theorien verändern und entwickeln sich mit der Zeit.
	Item 3	Einige Vorstellungen in den Naturwissenschaften sind heute anders als das, was Naturwissenschaftler früher dachten.
	Item 4	Naturwissenschaftliche Theorien werden verändert oder ersetzt, wenn neue Beweise vorliegen.
	Item 5	Manchmal verändern sich die Vorstellungen in den Naturwissenschaften.
	Item 6	Es gibt manche Fragen in den Naturwissenschaften, die auch Naturwissenschaftler nicht beantworten können.
	Item 7	Durch neue Entdeckungen kann sich verändern, was Naturwissenschaftler für richtig halten.
	Item 8	Die Vorstellungen in den Naturwissenschaften verändern sich manchmal.
Rechtfertigung	Item 1	Es ist wichtig, Experimente mehr als einmal durchzuführen, um Ergebnisse abzusichern.
	Item 2	Wenn Naturwissenschaftler Experimente durchführen, legen sie im Voraus einige Aspekte der Untersuchung fest.
	Item 3	Es ist wichtig, eine konkrete Vorstellung zu haben, bevor man mit einem Experiment beginnt.
	Item 4	Die Ideen zu naturwissenschaftlichen Experimenten kommen daher, dass man neugierig ist und darüber nachdenkt, wie etwas funktioniert.
	Item 5	Ein Experiment ist ein guter Weg um herauszufinden, ob etwas wahr ist.
	Item 6	Gute Theorien stützen sich auf die Ergebnisse aus vielen verschiedenen Experimenten.
	Item 7	In den Naturwissenschaften kann es mehrere Wege geben, um Vorstellungen zu überprüfen.

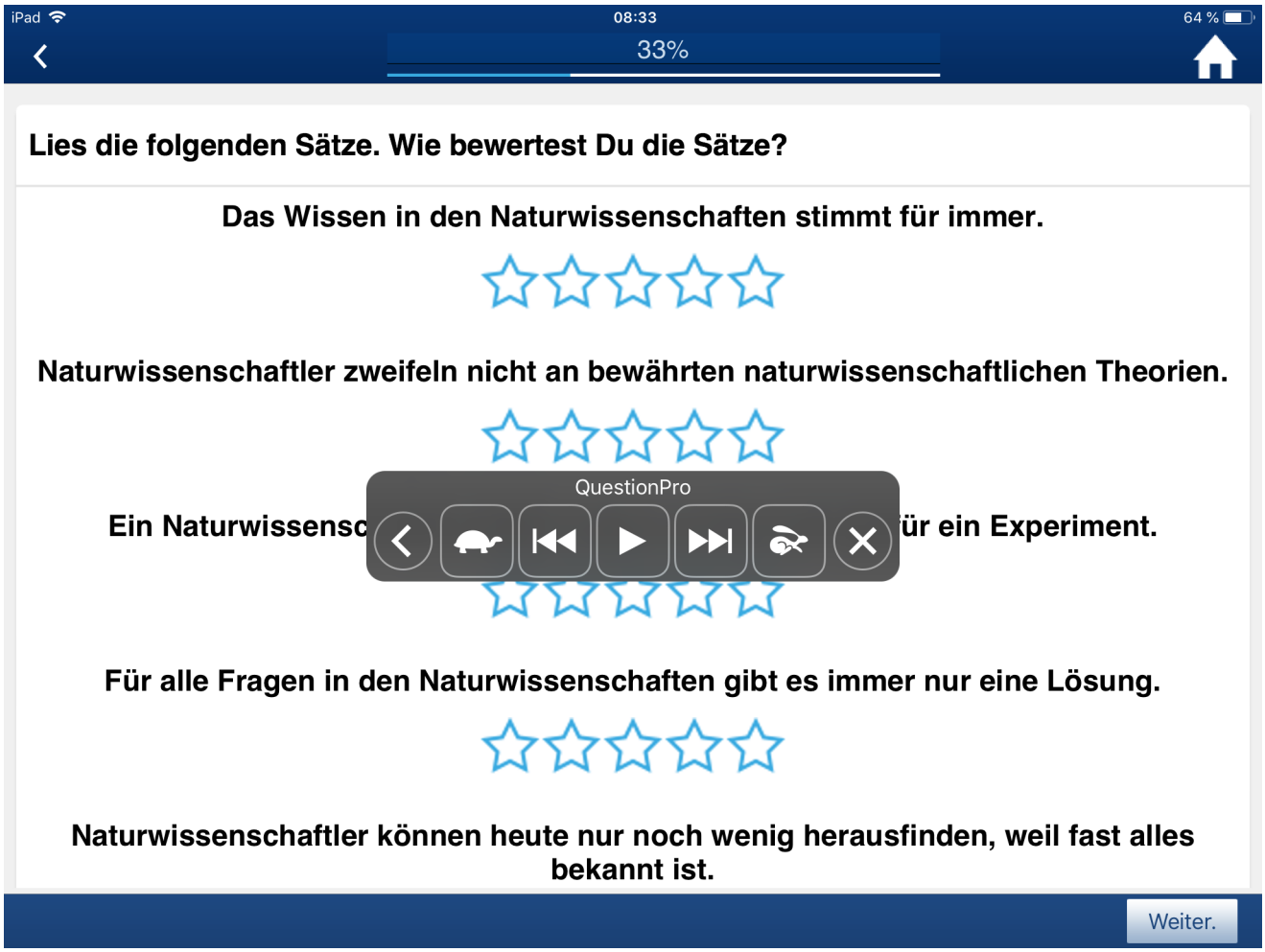


Abbildung 4.5: Abbildung des UDA-Assessments mit der implementierten Vorlesefunktion.

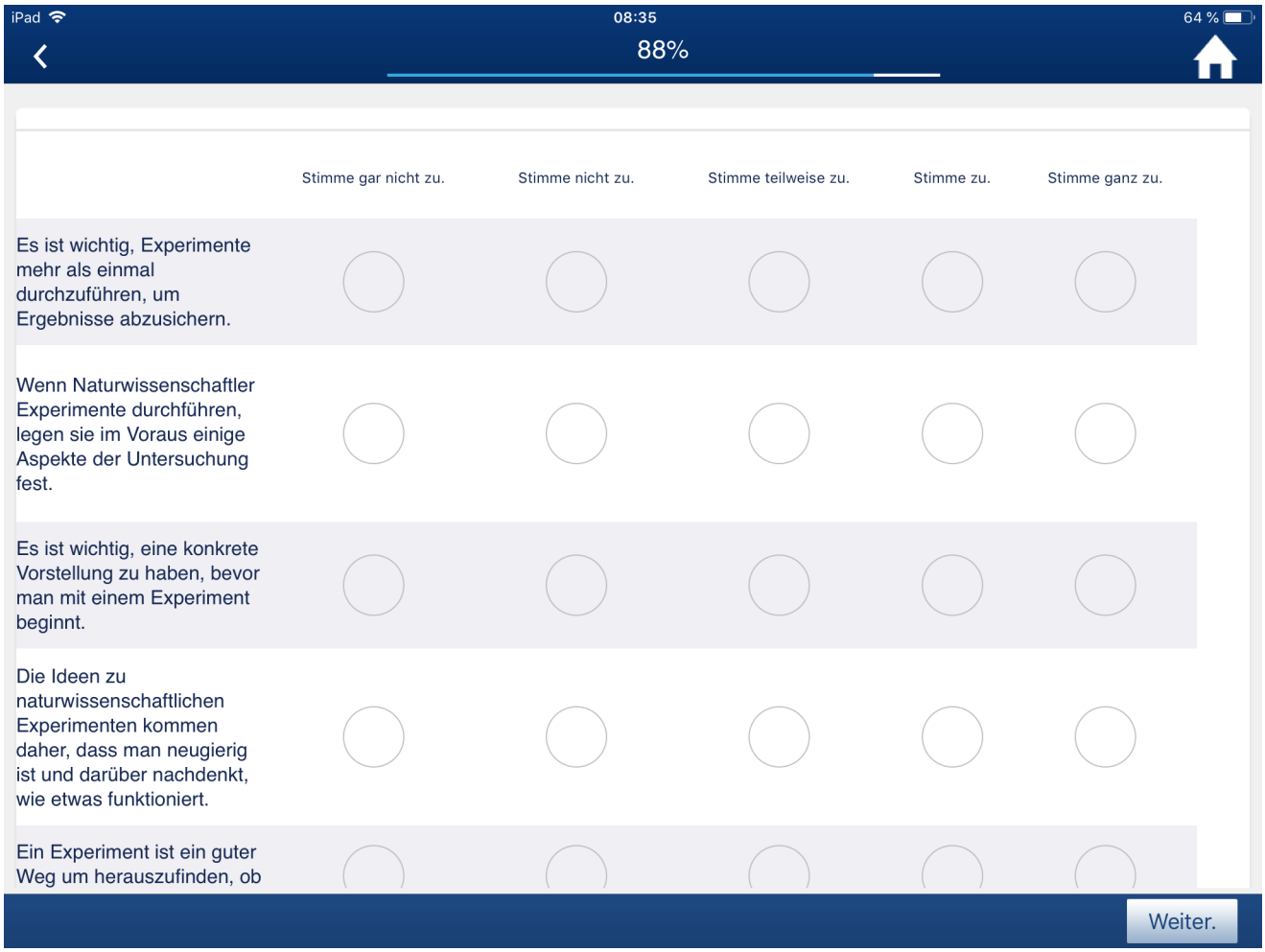


Abbildung 4.6: Abbildung des Originalassessments.

4.4.2 Die erhobenen Lernendenmerkmale

Für die Auswertung wurden weitere Lernendenmerkmale über Papier-Bleistift-Tests sowie über das iPad erhoben (Tab. 4.5). Die Auswahl erfolgte theoriebasiert. Im Sinne eines weiten Inklusionsverständnisses, was dieser Arbeit zu Grunde liegt, ist es nicht ausreichend, die Ergebnisse ausschließlich vor dem Hintergrund des sonderpädagogischen Unterstützungsbedarfs zu diskutieren. Zusätzlich wurde die Intelligenz, das Geschlecht, das Alter und ein diagnostizierter Förderbedarf erhoben. So sind vor allem die Sprachfähigkeit und der sozioökonomische Status von großer Bedeutung für den Bildungserfolg (Autorengruppe Bildungsberichterstattung, 2016, 2018).

Ausgehend von den Kriterien eines diversitätswertschätzenden Unterrichts (Tab. 2.1) wurden Daten zum wahrgenommenen Nutzen des E-Books (Williamson Sprague & Dahl, 2010) und der kognitiven Aktivierung erhoben (Fauth, Decristan, Rieser, Klieme & Büttner, 2014).

Tabelle 4.5: Erhobene Lernendemerkmale in dieser Studie.

Art des Tests	Konstrukt
Analoge Abfrage	Lesen: Salzburger-Lesescreening 2-9 (Mayringer & Wimmer, 2014) Kognitive Fähigkeiten: KFT 4-12+R-N2 (Heller & Perleth, 2000)
Digitale Abfrage	Sozioökonomischer Status: Torsheim et al., (2016) Kognitive Aktivierung: Fauth (2014) Wahrnehmung des Lernerfolgs: Sprague & Dahl (2009)
	Geschlecht Alter Diagnostizierter Förderbedarf

4.5 Statistische Methoden zur Datenauswertung der Hauptstudie

Für die Auswertung stellt sich auch die Frage nach statistischen Methoden zur Untersuchung der Wirkung der Maßnahmen zum Vergleich von Daten aus adaptierten und unveränderten Testinstrumenten und die Frage nach möglichst präzisen Skalen.

4.5.1 Bestimmung der longitudinalen Messinvarianz und der longitudinalen Elaborierung

Um die Vergleichbarkeit der Datenstrukturen aus dem 2x2-Between-Subject-Design zu bestimmen, können konfirmatorische Faktorenanalysen (KFA) verwendet werden. So kann die Gleichwertigkeit von Assessmentversionen longitudinal und mit Gruppenbezug untersucht werden (Brown, 2015; Kline, 2016; Little, 2013; Lovett & Lewandowski, 2015; Meade & Lautenschlager, 2004). Hierfür wird eine Modellstruktur spezifiziert. Anschließend wird über verschiedene Schätzmethoden überprüft, ob die Kovarianzstruktur der Daten zur Modellstruktur passt bzw. der Grad der Abweichung bestimmt.

Die Verwendung von Strukturgleichungsmodellierungstechniken (hierzu zählen KFAen) weist eine ganze Reihe von Vorteilen auf. Ein erster Aspekt betrifft die Frage nach der *Validität* des Messinstruments, da oftmals nicht direkt beobachtbare Konstrukte - zu denen die Konzepte von NOS zählen - untersucht werden. Strukturgleichungsmodelle erlauben den Einsatz einer Reihe von Datenformen, um das Zielkonstrukt zu beschreiben. Deutlich wird dieser Vorteil insbesondere bei latenten Wachstums- oder Panelmodellen, bei denen das Testformat gänzlich geändert werden kann, um Entwicklungen (z.B. von Intelligenz) über große Zeitspannen zu verfolgen (Little, 2013; Wu, Liu, Gadermann & Zumbo, 2010). Hierfür werden die Daten der Testinventare z-standardisiert.

Mit Blick auf die *Reliabilität* sehen sich statistische Methoden mit zum Teil erheblichen Anteilen von Messfehlern konfrontiert. Diese werden unter anderem durch Raten bei Testaufgaben verursacht. Erhöhte Messfehleranteile vermindern die Reliabilität von Variablen und schwächen damit die Beziehungen zwischen den Variablen. In der Folge werden bei-

spielsweise Korrelationen unterschätzt. In Strukturgleichungsmodellen können Messfehlereinflüsse in Variablen berücksichtigt werden. Schätzungen zu den Beziehungen zwischen dem/den Konstrukt/en sind dann unbeeinträchtigt von zufälligen Messfehleranteilen.

Statistisch relevant ist auch die Komplexität von Modellen. Strukturgleichungsmodelle beziehen meist viele Konstrukte gleichzeitig ein. So können die Beziehungen der Konstrukte in einem Gesamtmodell geschätzt werden. Hierbei können sowohl direkte als auch indirekte Effekte von Konstrukten aufeinander oder über Dritte bestimmt werden. Trotz der zum Teil sehr komplexen Modelle, erfüllen Strukturgleichungsmodelle das Prinzip der Sparsamkeit (Hoffmann, Minkin & Carpenter, 1997). Zur überprüfung von vergleichbaren Datenstrukturen wird die Messinvarianz (MI) der Daten aus den Gruppen und zu beiden Messzeitpunkten untersucht. Entsprechend müssen longitudinale KFAn vorgenommen werden, um interindividuelle änderungen zu bestimmen (Little, 2013; Newsom, 2015). Das Messmodell einer longitudinalen KFA weist sowohl korrelierte Varianzen der Indikatoren auf, als auch die Kovarianz der Konstrukte zu den jeweiligen Messzeitpunkten (Little, 2013) (Abb. 4.7). Darüber hinaus bedarf es einer Skalierung der Items über alle Messzeitpunkte. Dies kann entweder über den *common-factor*-Ansatz oder über die *effects-coded*-Methode erfolgen (Grimm, Ram & Estabrook, 2016; Little, 2013).

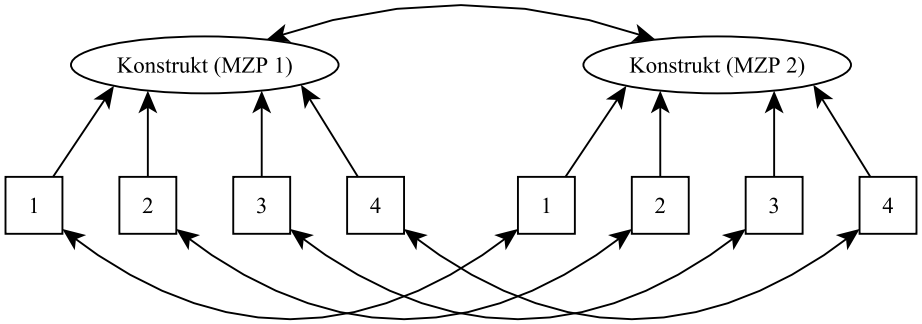


Abbildung 4.7: Vereinfachtes Messmodell einer longitudinalen KFA.

Je nach Grad der MI, können so Faktorstrukturen, Mittelwerte oder auch Fehlervarianzen verglichen werden. Wenn keine MI vorliegt, kann kein Vergleich der Testwerte aus den Gruppen vorgenommen werden (Kline, 2016) (Abschnitt 4.2). Aufgrund der kleinen Stichprobe in dieser Studie werden die KFA im *Multiple Indicator, Multiple Cause*-(MIMIC)- bzw. im *Multiple Group*-(MG)-Ansatz (Brown, 2015; Kline, 2016) und in der Kombination beider geschätzt (Sideridis, Tsaousis & Al-harbi, 2015). Während der MIMIC-Ansatz eine Kovariate (hier die Lernumgebung) in das Modell aufnimmt und auf den latenten Faktor sowie auf die manifesten Items regressieren lässt, vergleicht der MG-Ansatz zwei Kovarianzstrukturen von zwei oder mehr Substichproben vor einem gemeinsamen Strukturgleichungs- oder Messmodell.

Um die MI entsprechend der Gruppen und Messzeitpunkte zu prüfen, wird der Step-Up-Ansatz verwendet (Stark, Chernyshenko & Drasgow, 2006). Dieser sieht einen schrittweisen Vergleich weniger restriktiver mit restriktiveren Modellen vor. Die *konfigurale MI* ist die am wenigsten restriktive Form der MI. Hier sind sowohl die Anzahl der latenten Faktoren als auch die übereinstimmung zwischen den Faktoren gleichermaßen zwischen den Gruppen beschränkt. Alle anderen Parameter werden frei geschätzt. *Konfigurale MI*

kann angenommen werden, wenn grundsätzlich das gleiche Faktorladungsmuster zwischen Gruppen oder Messzeitpunkten vorliegt.Die nächste Ebene wird als schwache oder *metrische MI* bezeichnet. *Metrische MI* nimmt die Gleichheit aller Faktorladungen zwischen Gruppen oder Messzeitpunkten an. Liegt *metrische MI* vor, wird das gleiche Konstrukt in den Bedingungen bemessen. Wird neben der *metrischen MI* auch eine starke oder *skalare MI* von den Daten unterstützt, liegen vergleichbare Mittelwerte über die Gruppen vor. *Strikte MI* ist die restriktivste Form der Messinvarianz, weil die Gleichheit der Varianzen in allen Bedingungen unterstützt werden muss. Infolgedessen messen Indikatoren die Faktoren in jeder Gruppe oder zu jedem Messzeitpunkt mit der gleichen Genauigkeit. Diese Form der MI ist nicht unumstritten. Während Deshon (2004) und Wu, Zhen & Zumbo (2007) argumentieren, dass *strikte MI* erforderlich ist, um identische Messungen über Bedingungen hinweg durchzuführen, hält Little (2013) die Gleichsetzung von systematischen und zufälligen Messfehlern über die Zeit oder die Bedingungen hinweg für unangemessen.

Es kann der Fall sein, dass sich Modellparameter als ungeeignet zur Schätzung der Kovarianzstruktur erweisen. In diesem Fall sind diese Modellparameter messvariant bezüglich des Messzeitpunkts, der Lernumgebung oder der Testversion. über den Lagrange-Multiplier-Test kann ermittelt werden (Bentler & Chou, 1992), ob eine Aufhebung des messvarianten Modellparameters (z.B. die Gleichheit der Faktorladungen eines Items über alle Gruppen und Messzeitpunkte) zu einer Unterstützung der getesteten MI-Form führt. Sollte dies der Fall sein, liegt partiale MI auf der jeweiligen Stufe vor (Grimm et al., 2016; Little, 2013; Newsom, 2015). Während eine Messvarianz auf Ebene der Faktorladungen (*konfigurale und metrische MI*) für eine deutliche Heterogenität in den Daten für die Gruppen bzw. Messzeitpunkte spricht, stellt eine Messvarianz auf Ebene der Mittelwerte (*skalare MI*) lediglich eine unterschiedliche Entwicklung dar. Kann dennoch partial-skalare MI bestätigt werden, sind weiterhin die Konstrukte und die Mittelwerte miteinander vergleichbar.

Zur Durchführung der Analysen wird das R-Paket `lavaan` genutzt (Rosseel, 2012). Als Qualitätskriterien werden im ersten Schritt die üblichen Werte herangezogen (CFI ≥ 0.95, RMSEA ≤ 0.08, SRMR ≤ 0.08; (Hu & Bentler, 1999)). Für Longitudinale KFAen schlägt Little (2013) auch weniger restriktivere Evaluationskriterien vor (z.B.: CFI > 0.90) und begründet dies mit der Tatsache, dass mehr Beschränkungen in den Modelle vorgenommen werden und dass mehr Daten zur selben Probandin/zum selben Probanden vorliegen. Als Schätzer werden Maximum-Likelihood-Methoden (ML) verwendet, obwohl Likertskalen verwendet wurden (Liu et al., 2017). Dies wird durch die erhöhte Anzahl von fünf oder mehr Antwortkategorien möglich. So zeigten Rhemtulla, Brosseau-Liard & Savalei (2012), dass unter bestimmten Umständen ordinalskalierte Daten als kontinuierliche Variablen behandelt werden können. Dies gilt immer dann, wenn keine extreme Schwellenasymmetrie vorliegt. Außerdem sind für den ML-Schätzer auch robuste Varianten verfügbar, die die Schätzung von fehlenden Werten (unter der Annahme von "Missing at Random") über den Full-Information-Maximum-Likelihood-Algorithmus (FIML) erlauben. Letztlich weist der ML-Schätzer eine hohe Effizienz, gerade auch für kleine Stichproben auf (Urban & Mayerl, 2014).

Als Identifikationsansatz wird typischerweise die *Referentenmethode* im Feld der Strukturgleichungsmodellierung verwendet. Als Erweiterung dieser kann die *common-factor*-Methode für longitudinale Studien gesehen werden (Grimm et al., 2016). Der Referent dient dabei als Kalibriermaß, an dem alle anderen Items hinsichtlich ihres Betrages zur Varianzaufklärung des latenten Faktors ausgerichtet werden. Dies ist im Feld von NOS problematisch, weil aus einer theoretischen Perspektive der Referent als wichtigstes Item

einer Skala gelten sollte (Little, 2013; Newsom, 2015). Ein weiteres Problem ergibt sich, wenn gerade der Referent sich als messvariant herausstellt. Um mit diesem Problem umzugehen wird daher nicht der Identifikationsansatz über einen Referenten angewendet, sondern über die Skalierung der Faktorvarianz. Dieser *effects coding*-Ansatz nimmt an, dass sich der Mittelwert von n-Faktorladungen aus der Summe der geschätzten Werte der Faktorladungen geteilt durch die Summe der n-Faktorladungen ergibt.

$$\bar{\lambda} = \frac{\lambda_1 + ... + \lambda_n}{n}$$

Für ein Beispiel mit drei Indikatoren ergibt sich:

$$\lambda_{1,1} = 3 - \lambda_{2,1} - \lambda_{3,1}$$

Mit dieser Anforderung für die erste Faktorladung, wird die Faktorvarianz zu einer gewichteten Funktion der Kovarianz für alle Indikatoren (Newsom, 2015).

Um die *interindividuelle änderungen* zu den NOS-Skalen zu bestimmen, werden longitudinale Panelmodelle verwendet. In einem Panelmodell werden die Daten aller Probanden und Probandinnen gemittelt. Diese unterscheiden sich von longitudinalen KFA-Modellen im Wesentlichen dadurch, dass die Kovarianzen der latenten Konstrukte im Messmodell durch Regressionen ersetzt werden (bildlich gesprochen: Linien mit zwei Pfeilspitzen werden zu einem gerichteten Pfeil). Für die Analyse ergibt sich damit, dass die Werte der NOS-Skala zum ersten Messzeitpunkt die unabhängige und entsprechend die Werte derselben Skala zum zweiten Messzeitpunkt die abhängige Variable darstellt. Alternativ kann auch von der Prädiktor- (Werte der Skala zum Messzeitpunkt 1) und der Zielvariablen (Werte der Skala zum Messzeitpunkt 2) gesprochen werden. Außerdem kann der Einfluss von Kovariaten auf die *interindividuelle änderung* bestimmt werden (Little, 2013). Longitudinale Panelmodelle setzen zumindest partial-skalare MI voraus (Little, 2013).

Vor dem Hintergrund eines *Lehrens und Lernens für alle* werden latente Wachstumsmodelle zur Beschreibung der *intraindividuellen änderung* geschätzt (Grimm et al., 2016; Little, 2013; Wu et al., 2010). Latente Wachstumsmodelle beschreiben als Second-order-Modelle die Varianz der longitudinalen Mittelwerte als auch deren änderung zu den jeweiligen Messzeitpunkten (Wu et al., 2010) und setzen ebenfalls longitudinale KFA-Modelle voraus, die zumindest partial-skalare MI unterstützen. Mit ihnen ist es möglich, die änderung je Testteilnehmenden zwischen zwei oder mehr Messzeitpunkten in Abhängigkeit zu allen Probandinnen und Probanden der Stichprobe zu bestimmen. Folglich werden *individuelle Trajektorien* bestimmt. Diese variieren im ermittelten Wert zu jedem Messzeitpunkt und in der Steigung zwischen mindestens zwei Messzeitpunkten. Auch für diese Modelle wurde der *effect-coded*-Ansatz zum Skalieren der Mittelwerte und der Faktorladungen verwendet. Für das Messmodell werden die Mittelwerte der First-order-Konstrukte auf null fixiert ebenso wie deren Kovarianz. Die Varianz der First-order-Konstrukte wird über die Messzeitpunkte beschränkt. Die Faktorladungen für den longitudinalen Mittelwert wird über alle Messzeitpunkte frei geschätzt. Die änderungsrate muss mindestens einmal auf null fixiert und frei geschätzt werden. Je nach Studiendesign kann die Faktorladung dazwischen oder darüber hinaus frei geschätzt werden. Ein exemplarisches Messmodell für die Studie zeigt Abb. 4.8. Die Kovarianz der Second-order-Konstrukte kann frei oder fixiert geschätzt werden. Die Analyse wird auch in diesem Fall im Gruppenvergleich zu den Testversionen vorgenommen. Aus dem latenten Wachstumsmodell heraus können individuelle Trajektorien bestimmt und mit Hilfe von weiteren Variablen kontrolliert werden.

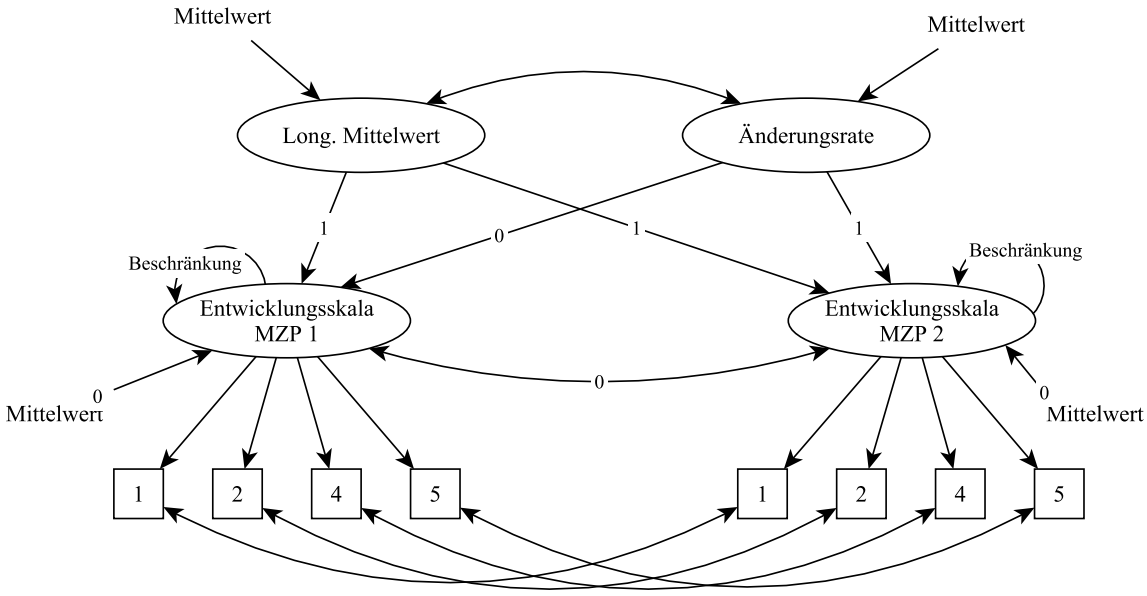


Abbildung 4.8: Vereinfachtes Modell eines latenten Wachstumsmodells. Eine freie Schätzung wird durch 1 angezeigt.

Mit latenten Wachstumsmodellen lässt sich die Kovarianz zwischen der änderungsrate und deren Richtung sowie der longitudinalen Mittelwerte bestimmen (Abb. 4.8). Die änderungen können dann wiederum mit Hilfe von Prädiktoren kontrolliert werden. Latente Wachstumsmodelle stellen insofern eine Erweiterung von longitudinalen KFA-Modellen dar.

4.5.2 Bestimmung der instruktional sensitiven Items

Ein weiterer wichtiger Schritt in der Auswertung besteht in der Bestimmung der *instruktionalen Sensitivität* der Assessmentformen. Die Suche nach instruktional sensitiven Items beeinflusst die Skalenbildung der Daten aus dem Assessment. Polikoff (2010) fast die Bedeutung der instruktionalen Sensitivität zusammen:

> The term 'instructional sensitivity' was chosen as the focal term for this analysis to refer to the extent to which student performance on a test or item reflects the instruction received (Kosecoff & Klein, 1974). [...] Furthermore, wherever the term instructional sensitivity' has been used, it has been to refer to a property of a test or item. In short, an instructionally sensitive test should be able to detect differences in instruction received by students" (Polikoff, 2010, S. 3).

Folglich kann ein Assessment, was zu viele Items umfasst, die nicht instruktionalsensitiv sind, zu einer Verschleierung von Effekten führen, wenn aus diesen Skalen Mittelwerte gebildet werden. Insbesondere mit Blick auf die Problematik, dass im Feld von UDL bisher kaum primäre Bildungseffekte beobachtet werden konnten (Capp, 2017), gilt diesem Aspekt in der Analyse eine besondere Aufmerksamkeit.

Es können verschiedenste Verfahren zur Bestimmung einer möglichen instruktionalen Sensitivität von Items genutzt werden. Neben den Items, die sich möglicherweise im Multigroup-Ansatz als messvariant erwiesen haben, können nach Brown (2015) auch Items auf ihre instruktionale Sensitivität im MIMIC-Ansatz geprüft werden. Instruktionale Sensitivität liegt dann vor, wenn die Regression der Kovariate auf ein Item signifikant ist.

Im Gegensatz hierzu nutzen Deutscher & Winther (2018) Differential Item Functioning (DIF) Analysetechniken mit den Messzeitpunkten als Gruppeneinteilung, während Naumann, Hartig & Hochweber (2017) Multilevel-IRT-Analysen vorschlagen. Es werden aber auch statistisch weniger anspruchsvolle Methoden diskutiert (Polikoff, 2010). Hierunter fallen z.B. T-Tests. So kann aber auch der prozentuale Anteil von richtig gelösten Items zum Prä- und Postmesszeitpunkt bestimmt werden. Außerdem können explorative Faktoranalysen vorgenommen werden, um Items zu identifizieren, die schlecht oder mäßig mit anderen einen gemeinsamen Globalfaktor bilden. Ist dies der Fall, liegt auch hier ein Indiz für instruktionale Sensitivität vor.

Für dieses Projekt ist die Frage nach der *instruktionalen Sensitivität* durch den expliziten Fokus der Lernumgebung auf die *NOS-Konzepte* aus dem Bereich *Rechtfertigung* von besonderer Bedeutung. So besteht die Frage, ob die Anlage der Lernumgebungen sich auch im Antwortverhalten der Testeilnehmer zum zweiten Messzeitpunkt widerspiegelt. Nur instruktional sensitive Tests können Unterschiede zu den eingesetzten Lernumgebungen aufzeigen. Entsprechend müssen die Skalen für die weiterführende Analyse ggf. gekürzt werden. Die Bestimmung einer möglichen instruktionalen Sensitivität der Items erfolgt über den latenten MIMIC-Ansatz und den manifesten Ansatz über T-Tests.

Eine weitere Möglichkeit die *instruktionale Sensitivität* von Skalen zu betrachten, liegt in der Schätzung von standardisierten Faktorladungen und somit in der Analyse der internen Validität einer Skala.

> "Aus der standardisierten Faktorladung lässt sich auch die gemeinsame Varianz von Faktor und Indikator berechnen. Sie entspricht dem Quadrat der standardisierten Faktorladung [...]. Die Höhe der standardisierten Faktorladung gilt auch als Hinweis auf die formale Validität (oftmals auch als *interne Validität* bezeichnet) eines Faktors" (Urban & Mayerl, 2014, S. 54, Herv. im Original).

Als Daumenregel gilt: Es sollte eine standardisierte Faktorladung von $>.50$, besser jedoch von $>.70$ vorliegen (Urban & Mayerl, 2014).

4.5.3 Bestimmung der Testzugänglichkeit

Darüber hinaus ergibt sich aus den Forschungsfragen (Abschnitt 3.2) die Anforderung, die Interaktion zwischen den Probandencharakteristiken und den Assessmentformen zu überprüfen. Um nicht nur die Vergleichbarkeit der Messinstrumente zu untersuchen, ist auch eine Untersuchung des Einflusses der Lernendenmerkmale (Geschlecht, Intelligenz, Lesefähigkeit oder sozioökonomischer Status) erforderlich. Zu diesem Zweck erweisen sich DIF-Ansätze als geeignet. DIF-Analysen zielen auf die Bestimmung gruppenabhängiger Indikatoren (Items). Hierüber wird unterschiedliches Antwortverhalten von Gruppen zum gleichen Item auf Basis von Probandencharakteristiken beschrieben. Hierfür wird ein *polytomes Rasch-Modell* mit der entsprechenden Anzahl der zu schätzenden Parameter angege-

ben. Ein *Partial-Credit-Modell* ist für diese Situation geeignet (Masters, 1982). Zusätzlich zur Schwierigkeit des Items schätzt das Modell Schwellenwerte für jede der Antwortkategorien pro Item (z.B.: vier bei fünf Antwortkategorien).

Eine grafische Messinvarianzprüfung findet dann in einem Koordinatensystem statt, in dem beide Achsen die Itemschwierigkeiten der gegeneinander aufgetragenen Untergruppen anzeigen. Die Messvarianz eines Items liegt vor, wenn die Ellipse der 95% -Konfidenzintervalle die Diagonale nicht berührt (Schwab & Helm, 2015). Eine Skala weist genau dann *Testzugänglichkeit* auf, wenn kein Item eine Gruppenabhängigkeit aufweist. In diesem Fall sind die Probandencharakteristiken nicht maßgeblich für das Antwortverhalten. In der vorliegenden Studie werden polytome Items mit einem fünfstufigen Antwortmuster analysiert. Für die DIF-Analysen wird das R-Paket `pairwise` verwendet (Heine, 2017).

4.5.4 Ablauf der Datenanalyse

Nach einer grundlegenden statistischen Beschreibung der Daten, wird überprüft, ob grundsätzlich longitudinale Messinvarianz vorliegt. Die Analyse der Vollskalen ist notwendig, um zu überprüfen, ob die Konstrukte aus beiden Assessmentversionen über die Gruppen und die Messzeitpunkte vergleichbar sind. Aus dem Literaturfeld zu Strukturgleichungsmodellierungen ist zu entnehmen, dass in der Regel drei bis fünf Items pro latentem Konstrukt vorliegen sollten (Brown, 2015; Grimm et al., 2016; Little, 2013; Newsom, 2015). Eine überhöhte Anzahl von Items führt zu deutlich überidentifizierten Modellen, was in der Folge die Varianz auf manifester und auf latenter Ebene erhöht. Dies kann durch eine große Stichprobe ausgeglichen werden. Jedoch reduziert sich die Genauigkeit der Messung. Es wird daher empfohlen, die Anzahl der Freiheitsgerade (= Summe der verfügbaren Informationen abzüglich der Summe der zu schätzenden Modellparameter) möglichst gering zu halten (Brown, 2015; Newsom, 2015). Es wird aber auch empfohlen, dass die Summe gleich null sein sollte (just-identified-Modelle) (Little, 2013).

Um ein möglichst genaues Bild zu zeichnen, werden daher die Vollskalen auf instruktional sensitive Items vor dem Hintergrund der Lernumgebungen untersucht. In der Folge werden Kurzskalen mit den zu der Lernumgebung passenden Items analysiert. Es erfolgt eine erneute überprüfung der longitudinalen Messinvarianz, weil sich die Kovarianzstruktur durch die Abänderung der Skalen ebenfalls ändert. Liegt für die Skalen zumindest partial-skalare MI vor, können die weiteren Analysen zur *Testzugänglichkeit* (Abhängigkeit Personen- und Itemcharakteristika), dem *Differential Boost* (Wirkung der Testadaptionen) und der *inter- und intraindividuellen änderung* (Vergleich der Mittelwerte aus allen Testwerten sowie Vergleich der Trajektorien aller Testeilnehmenden) erfolgen. Den Ablauf der Analyse zeigt Abb. 4.9.

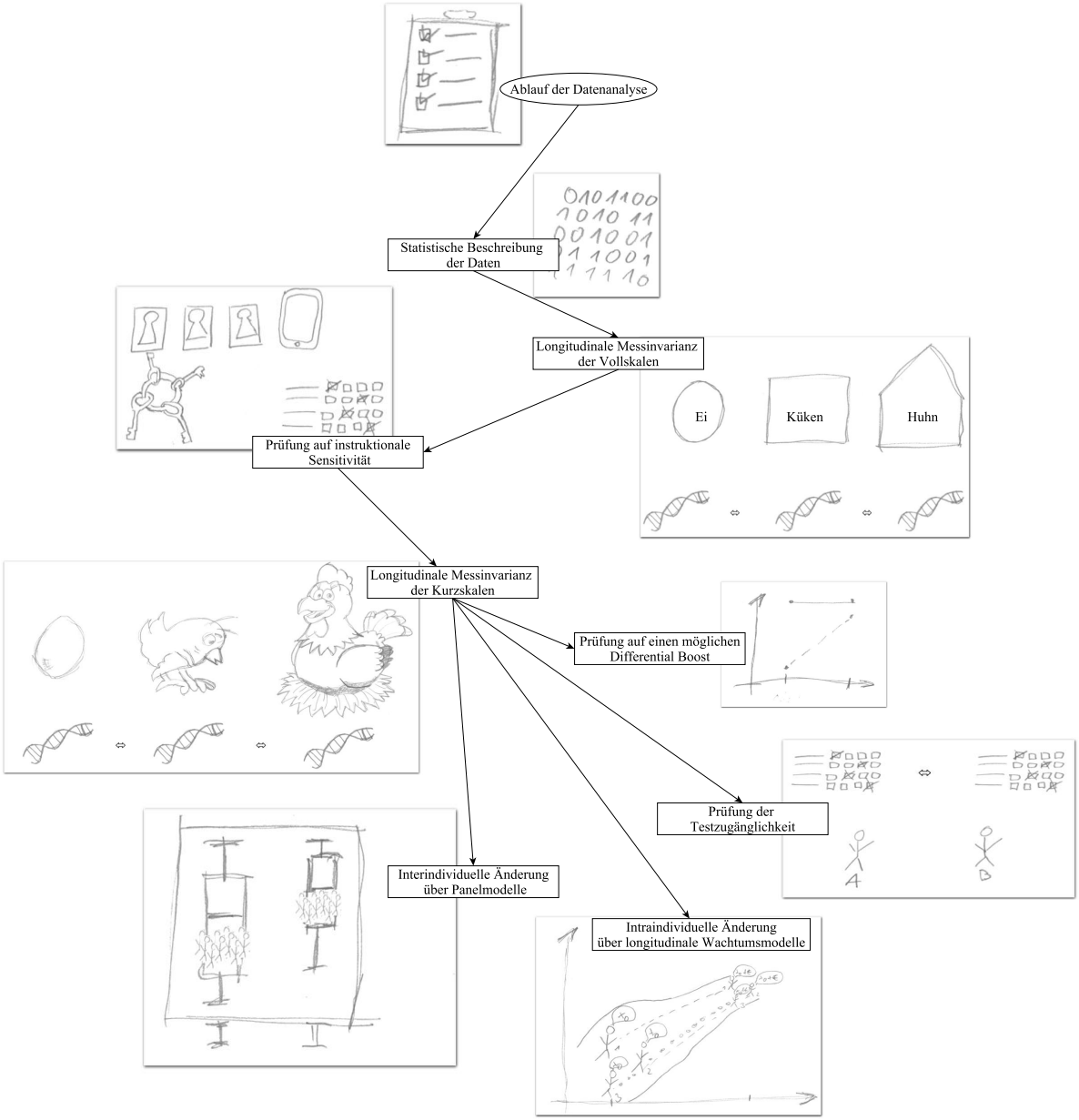


Abbildung 4.9: Grafische Darstellung der statistischen Auswertung.

5 | Ergebnisse der Studien

5.1 Vorstudienergebnisse

5.1.1 Beschreibung der Stichprobe und Datengewinnung in der Vorstudie

Die Stichprobe der Vorstudie umfasste 36 Lernende (männlich = 23 und weiblich = 13) aus der 5. bis 7. Klasse. Neben der Klassenstufe, wurde noch ein diagnostizierter sonderpädagogischen Unterstützungsbedarf als Hintergrundvariable erhoben. Aus der Stichprobe wiesen 9 Lernende zumindest einen sonderpädagogischen Unterstützungsbedarf auf. Die Schülerinnen und Schüler nahmen in Vierergruppen an den Vermittlungsversuchen teil und wurden im Vorfeld und im Nachgang einzeln leitfadengestützt interviewt (Abschnitt 7.5). Die Teilnehmer wurden symmetrisch auf die drei Lernumgebungen in Vierergruppen verteilt (12 Text, 12 Comic, 12 Video). Die Aussagen der Probandinnen und Probanden wurden mit Hilfe der Qualitativen Inhaltsanalyse (QI) kodiert (Kuckartz, 2012; Mayring, 2013).

Als Basis für die deduktive Analyse diente ein Kodiermanual von Carey et al. (1989) (Abschnitt 7.4). Als Hauptkategorien wurde das Niveau von Aussagen zum *Zweck des Experiments* und des *Prozesses beim Experimentieren* aufgenommen. Außerdem wurden deskriptive Aussagen zu den Lernzielen der Lernumgebung erfasst. Letztlich wurde aber auch induktiv vorgegangen, um dominante Themen oder Konzepte der Schülerinnen und Schüler zu berücksichtigen. Als kleinste Codiereinheit diente ein Halb- bzw. Nebensatz. Um die Intersubjektivität abzusichern, wurden 10 % des Materials doppelt kodiert. Hierfür wurden fünf Hauptkategorien verwendet. Die Interraterreliabilität in verschiedenen Maßen (Tab. 5.1) kann als gut bis sehr gut bezeichnet werden (Krippendorff, 2004).

Tabelle 5.1: Interraterreliabilität zur Absicherung der Intersubjektivität.

	übereinstimmung				
Doppelte Kodierungen	relativ	absolut	Cohens-k	Holsti	Krippendorfs-a
104	79 %	84	0.77	0.81	0.94

5.1.2 Vergleich der Themengebiete zu beiden Interviewzeitpunkten

In einem ersten Schritt wurden Aussagen aus den 72 Interviews gesammelt und zu Themengebieten systematisiert. Insgesamt lassen sich 167 Aussagen zu beiden Interviewzeitpunkten beobachten und in 13 Themengebiete zusammenfassen (Tab. 5.2). über alle Themengebiete hinweg ist auffällig, dass die Schülerinnen und Schüler in der Stichprobe noch keine Differenzierung zwischen NaturwissenschaftlerInnen und IngenieurInnen vornehmen. Diese werden undifferenziert als Erfinderinnen und Erfinder beschrieben, die Experimente nur beobachten und ohne spezifische Planung durchführen. Die Arbeit des Chemikers/der Chemikerin zum Interviewzeitpunkt 1 dient vor allem dem Ausprobieren (24.7 %, n = 22). Eng hiermit verwandt ist die Vorstellung, dass diese Stoffe mischen (3.4 %, n = 3) und auf einen Effekt hoffen. Chemikerinnen und Chemiker verfolgen mit ihrer Arbeit aber auch persönliche Zwecke und solche außerhalb des Naturwissenschaftlichen. So dient ihre Arbeit der Entdeckung von Dingen (16.9 %, n = 15), dem Menschenwohl (1.1 %, n = 1) oder dem persönlichen Erfolg (2.2 %, n = 2). Dabei ist die Arbeitsweise von Chemikerinnen und Chemikern vor allem auf Sicherheit ausgerichtet (10.1 %, n = 9), folgt aber nur einer unspezifischen Schrittfolge (6.7 %, n = 6).

Tabelle 5.2: Häufigkeiten der Themen je Interviewzeitpunkt.

	abs. und rel. Häufigkeiten der Themen	
Thema	Interviewzeitpunkt 1	Interviewzeitpunkt 2
Erfolg	2.2% (2)	0% (0)
Mischungsmetapher	3.4% (3)	0% (0)
Menschenwohl	1.1% (1)	5.1% (4)
Ausprobieren	24.7% (22)	0% (0)
Sonstiges	11.2% (10)	0% (0)
Sicherheit	10.1% (9)	0% (0)
Schrittfolge	6.7% (6)	23.1% (18)
Entdeckungen machen	16.9% (15)	10.3% (8)
Theoretische Konstruktion	9% (8)	3.8% (3)
Technologie	14.6% (13)	11.5% (9)
Planung	0% (0)	3.8% (3)
Ideenprüfung	0% (0)	41% (32)
Beobachtungscharakter	0% (0)	1.3% (1)

Zum Interviewzeitpunkt 2 sind die genannten Themen von der Arbeit mit der Lernumgebung beeinflusst (Tab. 5.2). So dient das Experiment der Prüfung von Ideen (41 %, n = 32) und weist Beobachtungscharakter auf (1.3 %, n = 1). Außerdem wird das Experiment geplant wozu eine bestimmte Schrittfolge benötigt wird (23.1 %, n = 18). Das Experiment dient weiterhin dazu, Entdeckungen zu machen (23.1 %, n = 18), dem Menschenwohl (5.1 %, n = 4) sowie der Technologieentwicklung (11.5 %, n = 9).

5.1.3 Vergleich der Aussagen zu den Lernzielen der Lernumgebungen

Alle drei Lernumgebungen (Video, Comic und Pop-Up-Text) fokussieren auf NOS-Konzepte zum *Zweck des Experiments* und zur *Planung um das Experiment*. Für die Analyse wurden Aussagen zum *Ideen testen* (hypothetisch-deduktive Arbeitsweise im Experiment (Labudde, 2010)) bzw. dem unsystematischen Ausprobieren (*Look & See*) (Carey et al., 1989) analysiert. Außerdem wurden Aussagen, die von einem "Konkreten" bzw. "unkonkreten Plan" bezüglich des experimentellen Vorgehens sprachen, gegensätzlich interpretiert und kodiert. Während zum ersten Interviewzeitpunkt lediglich 29 Aussagen in eine der vier Kategorien fielen, wurden zum zweiten Interviewzeitpunkt 51 mit Bezug zu einer dieser Kategorien geäußert. Die änderungen in der Verteilung sind zum Teil signifikant bestimmbar (Tab. 5.3). Das *Testen von Ideen* gewinnt zum zweiten Interviewzeitpunkt ebenso an Bedeutung, wie die *konkrete Planung* um das Experiment. Gegenläufig hierzu sind die Aussagen, die mit *Look & See* kategorisiert wurden. Lediglich das *unsystematische Planen* um das Experiment erfährt keine signifikante änderung. Insgesamt lassen sich auch fast doppelt so viele äußerungen im Postinterview einer der vier Kategorien zuordnen, wie im Präinterview.

Tabelle 5.3: Häufigkeiten der Kodierungen zu den Lernzielen je Interviewzeitpunkt.

Kategorie	Interviewzeitpunkt 1	Interviewzeitpunkt 2	Signifikanzniveau
Ideen testen	1	23	< 0.001
Look & See	23	8	< 0.01
Unkonkreter Plan	5	4	0.78
Konkreter Plan	0	16	< 0.001
Summierte Kodierungen	29	51	

In einem zweiten Schritt wurde diese Aussagen, den von Carey et al. (1989) entwickelten Niveaustufen zugeordnet (Tab. 5.4 und Tab. 5.5).

Tabelle 5.4: Kodierschema zum Zweck des Experiments.

Level	Beschreibung
Level 0	Lernende sehen den Zweck des Experiment nicht im Suchen von neuen Informationen.
Level 1	Lernende beschreiben den Chemiker/die Chemikerin, der/die ein Experiment dahingehend untersucht, ob es klappt. Die Motivation für das Experiment liegt nicht in der (Weiter-)Entwicklung von chemischen Wissen.
Level 2	Lernende unterscheiden zwischen der Testhypothese und dem Experiment an sich. Die Motivation für ein Experiment liegt in der überprüfung der Testhypothese.
Level 3	Lernende benennen den Zusammenhang zwischen den Ergebnissen eines Experiments und der Testhypothese. Die Ergebnisse dienen der Evaluation bzw. Entwicklung von chemischen Wissen.

Tabelle 5.5: Kodierschema zur Planung um das Experiment.

Level	Beschreibung
Level 0	Lernende beschreiben das Experiment als Ausprobieren.
Level 1	Lernende beschreiben das Experiment als unspezifisch geplanter Prozess.
Level 2	Lernende beschreiben das Experiment als spezifisch geplanter Prozess und nennen Schritte innerhalb des Prozesses.
Level 3	Zusätzlich zu Level 2 beschreiben die Lernenden die Abhängigkeit zwischen den Ergebnissen eines Experiments und dessen Design.

In Tabelle 5.6 sind beispielhafte Aussagen zu diesen Niveaustufen aufgeführt.

Tabelle 5.6: Beispielaussagen zu den Niveaustufen aus der Vorstudie.

Level	Zweck von Experimenten	Planung von Experimenten
0	Ich glaube sie vermischen halt mehrere chemische (ähm) (ähm) ja- Wie handelt man das? Halt mehrere Sachen so, dass halt verschiedene chemische Reaktionen so rauskommen (Abs. 16, S. 32).	Ja, wenn das gut ist, dann benutzen sie es. Wenn es schlecht ist, fangen sie von Neuem an (Abs. 1, S. 10).
1	Versuche rauszufinden. Also, wenn ja das mischt: Was dann passiert (Abs. 21 + 22, S. 33).	Sie probieren jetzt die Experimente aus und gucken, ob die Experimente gut sind, ob sie das auch weiter benutzen könnten und so (Abs. 39, S. 21)
2	Halt so wie gerade eben bei diesem Comic da wo die zwei sag ich mal irgendwie gestritten haben, dass (ähm) gleich viel auch gleich schwer ist und so, obwohl es gar nicht stimmte und als man das Experimente gemacht hat ist klargeworden, dass es nicht gleich schwer ist. Das man halt so zuerst so ein kleines Experiment machen sollte oder (Abs 23, S. 17).	Sie machen zuerst eine Vermutung oder schreiben auf ihre Vermutung, dann testen sie halt den Versuch und dann schreiben sie die (ähm)- was passiert ist auf (Abs. 1, S. 14).
3	Dann haben sie das jetzt ausprobiert. Und es war dann halt falsch. Also, es war falsch in dem Sinne, dass es nicht richtig war, dass (ähm) gleiches Volumen auch gleiches Gewicht [meint]. Es kommt auf die Materialien an (Abs. 9, S. 66).	Nicht vorgekommen.

Nicht nur die absolute Anzahl von äußerungen zu den Lernzielen der Lernumgebungen ändert sich. Um Aussagen zur Elaborierung der NOS-Konzepte zu treffen, wurde jeder und jedem der 36 Schülerinnen und Schüler zu beiden Zeitpunkten je ein Level für das gesamte Interview zugewiesen. Die Analyse vor dem Hintergrund der Niveaustufen zum *Zweck des Experiments* und zur *Planung um das Experiment* zeigt, dass die Verteilung der Niveaustufen sich durch die Intervention ändert (Tab. 5.7). Die Schülerinnen und Schüler argumentieren zu Beginn fast ausschließlich auf den Level 0 (n = 7 bzw. n = 27) und Level 1 (n = 27 bzw. n = 9) bezüglich beider Kategorien. Zum zweiten Interviewzeitpunkt ändert sich die Verteilung dahingehend, dass der Zweck von Experimenten vor allem im Testen von Hypothesen (Level 2: n = 22) besteht und die Daten auf diese bezogen werden müssen (Level 3: n = 22). Einige verbleiben aber auch bei der Vorstellung, dass Experimente lediglich klappen müssen (Level 1: n = 11). Folglich beschreiben die Schülerinnen und Schüler auch die Planung als (un-)systematische Abfolge von Teilschritten (Level 1: n = 12, Level 2: n = 15). Das Design und Ergebnisse miteinander in Verbindung stehen (Level 3), benennen die Teilnehmer und -nehmerinnen nicht. Diese änderungen für beide Kategorien sind höchst signifikant (Zweck des Experiments: Z = 0, $p < 0.001$; Prozess des Experimentierens: Z = 8, $p < 0.001$).

Tabelle 5.7: Niveauänderungen der Aussagen.

	Zweck des Experiments		Prozess des Experimentierens	
Level	Präinterview	Postinterview	Präinterview	Postinterview
Level 0	7	1	27	9
Level 1	27	11	9	12
Level 2	2	22	0	15
Level 3	0	2	0	0

Anmerkung:
Jedem Lernenden wurde ein Level je Interview zu gewiesen.

Die Beispielaussagen zeigen darüber hinaus auch die Tendenz der Studienteilnehmerinnen und -nehmer ihre Aussagen zum ersten Interviewzeitpunkt in der Regel ohne Bezug auf einen Kontext äußern (Deng, Chen, Tsai & Chai, 2011). Zum Interviewzeitpunkt 2 werden die Aussagen mit Bezug auf die Lernumgebung formuliert. Außerdem lassen sich nicht adressierte Vorstellungen auch nach der Bearbeitung der Lernumgebung wiederfinden. Hier finden oftmals Adaptionen statt. Allerdings werden auch nicht intendierte Vorstellungen im zweiten Interview benannt. So wird vor allem die Diskussion zwischen Wissenschaftlern als wichtiger Aspekt der wissenschaftlichen Arbeit beschrieben.

5.1.4 Analyse der Aussagen zu den Lernzielen der Lernumgebungen mit Bezug auf die Hintergrundvariablen

In einem letzten Schritt wurden die Niveaustufen der Aussagen der Schülerinnen und Schüler mit Bezug auf die erhobenen Hintergrundvariablen untersucht. Während zum ersten Interviewzeitpunkt sich keinerlei Abhängigkeiten zwischen den Aussagen und den Hintergrundvariablen beobachten lassen (Tab. 5.8), erweisen sich die Aussagen zum zweiten Interviewzeitpunkt zum *Zweck des Experiments* als signifikant zusammenhängend mit der

Klassenstufe und dem Vorhandensein eines *sonderpädagogischen Unterstützungsbedarfs*. Der Zusammenhang zeigt eine mittlerere Stärke.

Tabelle 5.8: Zusammenhangsanalyse zwischen den kodierten Niveaustufen und den Hintergrundvariablen zum ersten Interviewzeitpunkt.

Zusammenhang	š	Signifikanzniveau	Cramérs
Level zum Zweck des Experiments im Zusammenhang mit ...			
dem Geschlecht	1.28	0.72	0.19
der Lernumgebung	5.6	0.28	0.28
dem SU	2.46	0.25	0.26
der Klassenstufe	2.46	0.27	0.26
Level zur Planung um das Experiment im Zusammenhang mit ...			
dem Geschlecht	0.36	0.68	0.1
der Lernumgebung	3.56	0.22	0.31
dem SU	1.23	0.39	0.19
der Klassenstufe	1.23	0.4	0.19

Anmerkung:
SU: sonderpädagogischer Unterstützungsbedarf.

Tabelle 5.9: Zusammenhangsanalyse zwischen den kodierten Niveaustufen und den Hintergrundvariablen zum zwei Interviewzeitpunkt.

Zusammenhang	š	Signifikanzniveau	Cramérs
Level zum Zweck des Experiments im Zusammenhang mit ...			
dem Geschlecht	3.49	0.35	0.31
der Lernumgebung	8.18	0.19	0.34
dem SU	10.55	<0.05	0.54
der Klassenstufe	10.55	<0.05	0.54
Level zur Planung um das Experiment im Zusammenhang mit ...			
dem Geschlecht	0.36	0.83	0.1
der Lernumgebung	3.77	0.49	0.23
dem SU	0.53	0.89	0.12
der Klassenstufe	0.53	0.89	0.12

Anmerkung:
SU: sonderpädagogischer Unterstützungsbedarf.

Während sich der Zusammenhang zur *Klassenstufe* durch den überhang an Schülerinnen und Schülern aus der fünften Klasse erklären lässt (Tab. 5.11), erklärt sich der Zusammenhang zwischen dem *sonderpädagogischen Unterstützungsbedarf* und den Aussagen zum *Zweck des Experiments* durch die gleiche oder höherwertige Qualität von Aussagen von eben diesen Schülerinnen und Schülern (Tab. 5.10). Bezüglich der Kategorie *Planung um das Experiment* zeigen sich vergleichbare Verteilungen der Niveauzuordnungen.

Eine weitere Frage bestand hinsichtlich der Gleichwertigkeit der Lernumgebungen, die sich in der Art der Inhaltsrepräsentationsform unterscheiden. So kann keine Abhängigkeit zwischen diesen und den Aussagen zu beiden Interviewzeitpunkten der Schülerinnen

Tabelle 5.10: Verteilung der Niveaustufen in Abhängigkeit zum sonderpädagogischen Unterstützungsbedarf.

	Niveaustufen			
	0	1	2	3
Zweck des Experiments				
Kein SU	0	10	17	0
SU	1	1	5	2
Planung zum Experiment				
Kein SU	6	9	12	0
SU	3	3	3	0

Anmerkung:
SU: sonderpädagogischer Unterstützungsbedarf.

Tabelle 5.11: Verteilung der Niveaustufen in Abhängigkeit zur Klassenstufe.

	Niveaustufen			
	0	1	2	3
Zweck des Experiments				
5. Klasse	1	11	15	1
6. Klasse	0	0	3	1
7. Klasse	0	0	4	0
Planung zum Experiment				
5. Klasse	8	10	10	0
6. Klasse	0	2	2	0
7. Klasse	1	0	3	0

und Schüler beobachtet werden (Tab. 5.8 und Tab. 5.9). Außerdem lässt sich kein systematischer Einfluss der Inhaltsrepräsentationsform auf die Verteilung der Niveaustufen beobachten (Tab. 5.12). Es kann daher davon ausgegangen werden, dass die Lernumgebungen in keiner Weise Vorteile gegeneinander aufweisen.

5.1.5 Analyse der Aussagen Prä-Post-Vergleich

Für den Prä-Post-Vergleich muss sichergestellt werden, dass sich die Stichproben in Abhängigkeit zur eingesetzten Lernumgebung hinsichtlich der Varianz ausreichend zum ersten Interviewzeitpunkt ähneln. Diese Annahme wird durch eine non-parametrische Varianzanalyse (Kruskal-Wallis-Test) für beide Lernziele gestützt (*Zweck des Experiments*: $\chi^2(2) = 3.91$, $p =$ n.s.) und *Planung um das Experiment*: $\chi^2(2) = 3.46$, $p =$ n.s.) in Abhängigkeit zur Lernumgebung unterstützt. Es besteht kein signifikanter Einfluss auf die Verteilung der Kategorien. Entsprechend kann der Prä-Post-Vergleich vorgenommen werden, indem die Level zum ersten von denen zum zweiten Interviewzeitpunkt subtrahiert werden.

Bezieht man die kodierten Levels zu beiden Interviewzeitpunkten aufeinander, lassen sich Unterschiede bezüglich der Niveaustufenänderungen in Abhängigkeit zur Inhaltsrepräsen-

Tabelle 5.12: Verteilung der Niveaustufen in Abhängigkeit zur Lernumgebung im Prä-Post-Vergleich.

	Niveaustufen							
	Präinterview				Postinterview			
	0	1	2	3	0	1	2	3
Zweck des Experiments								
Comic	3	9	0	0	1	4	7	0
Text	0	11	1	0	0	6	5	1
Video	4	7	1	0	0	1	10	1
Planung zum Experiment								
Comic	11	1	0	0	4	3	5	0
Text	7	5	0	0	3	6	3	0
Video	9	3	0	0	2	3	7	0

tationsform beobachten (Tab. 5.13). So lassen sich die äußerungen der Schülerinnen und Schüler zum *Zweck des Experiments* zum zweiten Interviewzeitpunkt in eine höhere Niveaustufe einordnen (Text: n = 6, Comic: n = 7, Video: n = 7). Allerdings lässt sich auch eine ähnlichen Anzahl der Schülerinnen und Schüler beobachten, die mit der comic- oder textbasierten Lernumgebung gearbeitet haben und keine änderungen zeigen (Text: n = 6, Comic: n = 4, Video: n = 1).

Bezüglich der *Planung um das Experiment* zeigt sich ein einheitliches Bild zwischen der comic- und der videobasierten Lernumgebung (Tab. 5.13). Hier treffen ca. 2/3 der Schülerinnen und Schüler elaboriertere Aussagen zum zwei Interviewzeitpunkt (Comic (1 Stufe): n = 4, Comic (2 Stufen): n = 4; Video (1 Stufe): n = 4, Video (2 Stufen): n = 5). Auch nach der Arbeit mit der textbasierte Lernumgebung werden die Aussagen elaborierter (Text (1 Stufe): n = 6, (Text (2 Stufen): n = 1).

Tabelle 5.13: Verteilung der Niveaustufennänderungen in Abhängigkeit zur Lernumgebung im Prä-Post-Vergleich.

	Niveaustufennänderungen			
	-1	0	1	2
Zweck des Experiments				
Comic	0	4	7	1
Text	0	6	6	0
Video	0	1	7	4
Planung zu Experiment				
Comic	0	4	4	4
Text	1	4	6	1
Video	0	3	4	5

Eine zweite Zusammenhangsanalyse zum zweiten Interviewzeitpunkt (Tab. 5.14) unterstützt diesen Eindruck. Demnach hat die Lernumgebung einen signifikanten Einfluss auf das Niveau der Aussagen im Postinterview ($\chi^2 = 8.75$, $p < .1$, $\Phi_c = 0.35$). Dies gilt

Tabelle 5.14: Zusammenhangsanalyse zwischen den Niveaustufenänderungen und den Hintergrundvariablen.

Zusammenhang	Chi-Quadrat	p-Wert	Cramérs phi
Level zum Zweck des Experiments im Zusammenhang mit ...			
dem Geschlecht	0.77	0.71	0.15
der Lernumgebung	8.75	0.058	0.35
dem SU	2.35	0.42	0.26
der Klassenstufe	2.35	0.43	0.26
Level zur Planung um das Experiment im Zusammenhang mit ...			
dem Geschlecht	1.93	0.68	0.23
der Lernumgebung	5.35	0.51	0.27
dem SU	2.08	0.62	0.24
der Klassenstufe	2.08	0.61	0.24

Anmerkung:
SU: sonderpädagogischer Unterstützungsbedarf.

jedoch nicht für das *Geschlecht*, die *Klassenstufe* und den *sonderpädagogischen Unterstützungsbedarf*, die jeweils keinen signifikanten Zusammenhang zur Niveaustufenänderung aufweisen.

So zeigt eine Varianzanalyse zur Differenz der Niveaustufen zum *Zweck des Experiments* ($F(2{,}33) = 5.02$, $p < .05$), dass der Unterschied von der genutzten Lernumgebung abhängt. Dies gilt nicht für die Differenz der Niveaustufen zur *Planung um das Experiment* ($F(2{,}33) = 1.58$, $p =$ n.s.). Eine Post-hoc-Analyse mit Bonferroni-Korrektur zeigt darüber hinaus, dass sich ausschließlich die text- (M = 0.5, SD = 0.52) und die videobasierte Lernumgebung (M = 1.25, SD = 0.62) bezüglich der änderung der kodierten Level signifikant und mit geringer bis mittlerer Effektstärke voneinander unterscheiden ($\eta^2 = 0.23$, $p < .05$). Die übrigen Vergleiche (text- vs. comicbasiert & comic- vs. videobasiert) unterscheiden sich nicht signifikant.

Tab. 5.15 zeigt beispielhafte Aussagen einiger Studienteilnehmerinnen und -teilnehmer vor dem Hintergrund der Lernumgebung und des sonderpädagogischen Unterstützungsbedarfs. Insgesamt wird das Experiment zum zweiten Interviewzeitpunkt als Mittel der Beweisführung gegenüber einer Frage- oder Problemstellung betrachtet. Dabei wird es einschränkend jedoch noch mit einer *autoritären Beschreibung* (richtig gegen falsch) versehen. Durch die Veranlagung der Lernumgebung wird auch die fachliche Kommunikation zwischen Chemikerinnen und Chemikern von einigen Lernenden als Teilschritt des Experimentierens beschrieben. Insgesamt wird der Prozess des Experimentierens nach der Intervention als linear jedoch mit offenem Ausgang durch die Schülerinnen und Schüler verstanden.

Tabelle 5.15: Beispielhafte Aussagen im Prä-Post-Vergleich.

	Beispielaussagen	
Beschreibung	Präinterview	Postinterview
Schülerin ohne SU, 6. Klasse, videobasierte Lernumgebung	Ach Gott, ne keine Ahnung (Abs. 7, S. 10).	Sie überlegen sich erstmal was und haben eine Ideen und versuchen es mit Experimenten umzusetzen (Abs. 6, S. 15).
Schüler mit SU, 6. Klasse, videobasierte Lernumgebung	Versuche rauszufinden. Also, wenn man das mischt: Was dann passiert (Abs. 21+ 22, S. 33).	dann ihre- probieren gegenseitig ihre Ideen aus und (ähm) gucken, ob das so richtig, wie die Idee gesagt ist (Abs. 48, S. 65).
Schülerin ohne SU, 5. Klasse, comicbasierte Lernumgebung	Also ich- Erstmal schützen sie sich mit einer Schutzbrille und alles und dann gucken sie, ob (ähm) - Sie holen sich die ganzen Materialien und dann Experimentieren sie erst (Abs. 1, S. 7).	Sie experimentieren ihre Ideen. Zum Beispiel sie wissen ja nicht, ob ihre Idee zum Beispiel schon klappt. Deswegen Experimentieren sie erstmal. Und gucken- und experimentieren nach der Antwort. Also, die Antwort. Also, ob es stimmt- Also, ob es richtig ist. (Abs. 11 + 13, S. 61).
Schüler mit SU, 7. Klasse, textbasierte Lernumgebung	Sie haben meistens eine Schutzbrille oder so auf und so Gläser, wo sie dann Experimentieren (Abs. 16, S. 11).	Sie testen die Ideen, die sie haben (Abs. 29, S. 63).

Anmerkung:
SU: sonderpädagogischer Unterstützungsbedarf.

5.1.6 Zusammenfassung der Vorstudienergebnisse

Forschungsfrage 1: Inwiefern führen die Lernumgebungen zu unterschiedlichen Elaborierungen von NOS-Konzepten?

Bezüglich der ersten Forschungsfrage lassen die Ergebnisse den Schluss zu, dass die videobasierte Lernumgebung Vorteile gegenüber der text- aber nicht gegenüber der comicbasierten Lernumgebung aufweist. Dies gilt allerdings nur, wenn die Niveauänderungen und damit eine intraindividuelle änderung einbezogen wird. Die Zusammenhangsanalyse hingegen lässt keinen Schluss auf eine überlegenheit der videobasierten Lernumgebung zu, wenn lediglich die Verteilung der Niveaustufen in Abhängigkeit zur Lernumgebung untersucht wird. Zum Interviewzeitpunkt 1 lassen sich keine signifikanten Unterschiede der Substichproben beobachten.

Forschungsfrage 2: Inwiefern führen die entwickelten Lernumgebungen zu einer Elaborierung von NOS-Konzepten bei allen Schülerinnen und Schülern?

Die zweite Forschungsfrage ist für beide Kategorien und für alle Schülerinnen und Schüler zu verneinen. Zwar lassen sich Niveaustufenänderungen über alle Lernumgebungen beobachten (*Zweck des Experiments*: 69.4 % und *Planung um das Experiment*: 66.7 %). Jedoch halten auch einige Schülerinnen und Schüler an ihren Konzepten fest (*Zweck des Experiments*: 30.6 % und *Planung um das Experiment*: 30.6 %). Auffällig ist jedoch, dass die Probandinnen und Probanden kaum Aussagen auf Level 3 tätigen. Hierfür wären vermutlich weitere Lernumgebungen notwendig die explizit auf den Zusammenhang von Design, Datengewinnung und Interpretation fokussieren. Die Lernumgebung führt folglich zu einer Elaborierung zu den NOS-Konzepten zum *Zweck des Experiments* und zur *Planung um das Experiment*.

Forschungsfrage 3: Inwiefern beeinflusst der sonderpädagogischen Unterstützungsbedarf die Elaborierung von NOS-Konzepten im Vergleich zwischen den Schülerinnen und Schülern?

Mit Bezug zur dritten Forschungsfrage lässt sich feststellen, dass es keinen negativen Zusammenhang zwischen dem Vorhandensein eines *sonderpädagogischen Unterstützungsbedarfs* und der Elaborierung der adressierten NOS-Konzepte beobachten lässt (Tab. 5.9). Entsprechend werden Schülerinnen und Schüler mit *sonderpädagogischem Unterstützungsbedarf* nicht benachteiligt und können qualitativ gleichwertige, oder bessere Antworten im Vergleich zu ihren Mitschülerinnen und -schülern erzielen (Tab. 5.10). Und es lassen sich auch keine Hinweise für einen Schereneffekt beobachten, der eine Bevorteilung von Schülerinnen und Schülern ohne sonderpädagogischen Unterstützungsbedarf anzeigen würde.

5.2 Hauptstudienergebnisse

5.2.1 Beschreibung der Stichprobe und der Daten

Die Stichprobe der Hauptstudie umfasste 348 Lernende (männlich = 189 und weiblich = 193 mit einem durchschnittlichen Alter von 12.2(0.74) Jahren. Alle Schülerinnen und Schüler stammen von hannoverschen Integrierten Gesamtschulen (IGS), die zum Feld der allgemeinbildenden Schulen zählen. IGS'en wurde oftmals aus Hauptschulen heraus gerundet und setzten in Niedersachsen als erste Schulen Modelle für einen integrativen Unterricht um. Entsprechend haben die Lehrerinnen und Lehrer dieser Schulform bereits einige Erfahrung in der Beschulung von Lernenden mit sonderpädagogischem Unterstützungsbedarf. In der Stichprobe der Hauptstudie hatten 16 Lernende einen diagnostizierten sonderpädagogischen Unterstützungsbedarf (Lernen: n = 12, Sprache n = 4). Dies entspricht einem Anteil von 4.6 %. Damit ist die Stichprobe leicht über der Quote von Niedersachsen aus dem Schuljahr 2014/2015, die bei einem Anteil von 3.9 % an allgemeinbildenden Schulen lag (Werning & Thoms, 2017).

Tabelle 5.16: Verteilung der Stichprobe auf die vier Untersuchungsgruppen.

	UDL-Lernumgebung	MR-Lernumgebung
UDA-Assessment	96	79
Originalassessment	87	78

Die Daten unterstützen nicht die Annahme einer multivariaten Normalverteilung für die Daten aus beiden Assessmentformen zum ersten Messszeitpunkt (Marida-Test: UDA-Assessment: $\chi = 4636.28$, $p < 0.001$, $z = 5.64$, $p < 0.001$; Originalassessment: $\chi = 4812.52$, $p < 0.001$, $z = 8.61$, $p < 0.001$) wie zum zweiten Messzeitpunkt (UDA-Assessment: $\chi = 5111.67$, $p < 0.001$, $z = 8.85$, $p < 0.001$; Originalassessment: $\chi = 6713.77$, $p < 0.001$, $z = 23.46$, $p < 0.001$). Jedoch sind die Muster in den Antwortenkategorien pro Item nicht übermäßig asysmmetrisch verteilt (mehr als 90 % der Antworten fallen auf eine Kategorie) (Rhemtulla et al., 2012). Im UDA-Assessment liegt keine Kategorie mit einem prozentualen Anteil von mehr als 57 % (NOS-Skala Sicherheit, Item 6, Messzeitpunkt 1) vot (Tab. 7.1). Im Orginalassessment sind es in des lediglich 45 % (NOS-Skala Rechtferigung, Item 5, Messzeitpunkt 1), die prozentual in eine Kategorie fallen (Tab. 7.2). In keinem Fall wurden Outlier aus den Daten entfernt. Dies hat zwei Gründe. Bezüglich der Likertskalen ist diese als ganzheitlich zu sehen. In einem Kontinuum von "Stimme gar nicht zu" bis "Stimme vollständig zu" kann es keine Outlier geben. Auch die übrigen Variablen wurden nicht von Outliern bereinigt, um die Diversität der Lernenden vollständig abzubilden.

Eine Daumenregel zu fehlenden Werten besagt, dass diese 20 % nicht überschreiten sollten (Jamshidian, Jalal & Jansen, 2014; Little, 1988; Newsom, 2018). Zum ersten Messzeitpunkt liegen die fehlenden Werte bezüglich der NOS-Items in der UDA-Version bei 8.05 % und unterscheiden sich damit leicht vom Originalassessment mit 11.21 % (Tab. 7.10 und Tab. 7.11). Zum zweiten Messzeitpunkt steigt der Anteil von fehlenden Werten für alle NOS-Items und in beiden Assessmentversionen (UDA-Version: 11.08 %, Originalassessment: 15.60 %, Tab. 7.10 und Tab. 7.12).

Da mit Blick auf die fehlenden Werte und für die Nutzung des Full-Information-Maximium-Likelihood (FIML) Algorithmus zumindest die *Missing at random* (MAR) Annahme gegeben sein muss, schlagen Jamshidian & Jalal (2010) vor, Tests auf der Basis der Gleichheit von Kovarianzen zwischen Gruppen, die aus identischen fehlenden Datenmustern bestehen, durchzuführen. Diese Tests zielen auf die restriktivere Form der *Missing completely at random* (MCAR). über den Hawkinstest auf mutlivariaten Normalverteilung und Homoskedastizität wird ein parametrischer und non-parametrischer Test zur Gleichheit der Kovarianzen in den Gruppen durgeführt. Für die Durchführung dieser Tests wird das R-Paket `MissMech` verwendet (Jamshidian et al., 2014). Da keine multivariate Normalverteilung der Daten vorliegt, wird der non-parametrische Test für die Untersuchung der Kovarianzstrukturen verwendet.

Im UDA-Assessment weisen die *Herkunftsskala* drei, die *Sicherheitsskala* zwei, die *Entwicklungsskala* fünf und die *Rechtfertigungsskala* vier Gruppen auf, die sich jedoch hinsichtlich ihrer Kovarianzstruktur nicht signifikant voneinander unterscheiden. Im Originalassessment ergeben sich für die *Herkunftsskala* vier, für die Sicherheitsskala ergeben sich drei, für die *Entwicklungsskala* ergeben sich drei und für die *Rechtfertigungsskala* drei Gruppen, die sich ebenfalls nicht hinsichtlich ihrer Kovarianzstruktur signifikant voneinander unterscheiden. Damit unterstützen die Daten die MCAR-Annahme.

Insgesamt sind die Mittelwerte, Standardabweichungen, Schiefe- und Wölbungswerte für beide Assessmentformen (Tab. 7.3 und 7.4) ähnlich. Auffällig ist jedoch, dass die Standardabweichungen im UDA-Assessment größer sind. Das heißt, es liegt mehr Varianz in dem Antwortverhalten vor.

5.2.2 Bestimmung der Messinvarianz zwischen den Testversionen, den Lernumgebungen und den Messzeitpunkten

5.2.2.1 Interne Konsistenzen bei Vorgabe der Faktorstruktur

Wird die Faktorstruktur (longitudinal und mit τ-äquivalenz; *effects-coded*-Ansatz (Little, 2013; Newsom, 2015)) vorgegeben, erweisen sich die Skalen insgesamt als ausreichend konsistent zum ersten Messzeitpunkt, mit Ausnahme der *Rechtfertigungsskala* im Originalassessment (Tab. 5.17). Mit Blick auf das UDA-Assessment sind alle Skalen mit Ausnahme der *Herkunftsskala* konsistenter. Zum zweiten Messzeitpunkt steigen alle internen Konsistenzen an. Am deutlichsten steigt ω_g für die *Rechtfertigungsskala* im Originalassessment (Messzeitpunkt 1: 0.52; Messzeitpunkt 2: 0.77). Die Verschiebungen können zwei Gründe aufweisen. Zum einem kann ein Re-Test-Effekt vorliegen, d.h. der Zuwachs ist auf bloße Wiederholung des Tests zurückzuführen. Zum anderem kann auch die Intervention eine Wirkung aufweisen, so dass die Steigerung der internen Konsistenz Lernzuwächse repräsentieren. Insgesamt kann jedoch durch die guten und zum Teil konstanten internen Konsistenzen davon ausgegangen werden, dass die Skalen keine Substruktur aufweisen. Allerdings erweisen sich einige Indikatoren als weniger geeignet zur Erfassung des latenten Konstrukts.

Tabelle 5.17: Interne Konsistenzen (McDonalds-ω) bei vorgegebener Skalenstruktur.

NOS-Skala	Messzeitpunkt 1	Messzeitpunkt 2	Beide Messzeitpunkte
UDA-Assessment			
Herkunft	0.59	0.83	0.76
Sicherheit	0.81	0.87	0.87
Entwicklung	0.86	0.89	0.9
Rechtfertigung	0.83	0.84	0.87
Originalassessment			
Herkunft	0.7	0.76	0.79
Sicherheit	0.78	0.85	0.87
Entwicklung	0.83	0.9	0.89
Rechtfertigung	0.52	0.77	0.74

Tabelle 5.18: Messinvarianzmodelle zur NOS-Skala Herkunft.

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	103.43	58	<0.05	0.08	0.929	0.889	0.057	Ja
Metrisch	119.13	70	<0.05	0.076	0.923	0.901	0.07	Ja
Skalar	180.01	82	<0.05	0.099	0.846	0.831	0.089	Nein
Partial skalar	131.39	79	<0.05	0.074	0.918	0.906	0.072	Ja
Strikt	169.78	88	<0.05	0.087	0.872	0.869	0.086	Nein

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für die Mittelwerte von Item 3 aufgehoben
Angen.?: Angenommen?

5.2.2.2 Messinvarianzprüfung der einzelnen Vollskalen und Bestimmung der instruktionalen Sensitivität

Um Aussagen zur Vergleichbarkeit der vier Gruppen zu beiden Messzeitpunkten machen zu können, werden im Folgenden Messinvarianzprüfungen über longitudinale konfirmatorische Faktorenanalysen (KFA) vorgenommen.

Herkunftsskala

Die Messinvarianzprüfung der *Herkunftsskala* zeigt, dass mit weniger konservativen Evaluationskriterien für die longitudinale Messinvarianzprüfung partial-skalare Messinvarianz angenommen werden kann (CFI = 0.918; Tab. 5.18). Item 3 erweist sich in dieser Skala jedoch als messvariant über beide Assessmentformen zum zweiten Messzeitpunkt. Die Instrumente messen daher grundsätzlich das gleiche Konstrukt und haben gleiche Nullpunkte. Lediglich die Fehlergenauigkeit in der Messung unterscheidet sich.

Die MIMIC-Analyse lässt keinen Rückschluss auf eine Messinvarianz hinsichtlich der Items in der Herkunftsskala zu (Tab. 5.19). Keine der Regressionen ist signifikant und der CFI

Tabelle 5.19: Regressionen im MIMIC-Ansatz zur Herkuntfstskala. Dargestellt sind die Regressionen der Lernumgebung auf den latenten Faktoren und die manifesten Indikatoren in einem konfiguralen Messinvarianzmodell.

	UDA-Assessment		Originalssessment	
	Schätzung	p-Wert	Schätzung	p-Wert
Latenter Faktor	0.15	0.26	-0.17	0.07
Item 1	0.27	0.12	-0.21	0.27
Item 2	0.09	0.49	-0.03	0.85
Item 3	-0.14	0.36	0.05	0.68
Item 4	0.10	0.42	0.05	0.70
Item 5	-0.12	0.44	0.19	0.15

Anmerkung:
CFI: 0.911; TLI: 0.861; RMSEA: 0.075; SRMR: 0.067

Tabelle 5.20: Messinvarianzmodelle zur NOS-Skala Sicherheit.

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	253.86	138	<0.05	0.083	0.904	0.873	0.064	Ja
Metrisch	275.62	156	<0.05	0.079	0.901	0.884	0.073	Ja
Skalar	412.36	174	<0.05	0.106	0.803	0.793	0.103	Nein
Partial skalar	328.45	165	<0.05	0.09	0.865	0.851	0.086	Nein
Strikt	348.82	181	<0.05	0.087	0.861	0.86	0.089	Nein

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für die Mittelwerte von Item 3 aufgehoben
Angen.?: Angenommen?

(= 0.902) auf dieser Messinvarianzstufe weiterhin mit einem akzeptablen Wert.

Sicherheitsskala

Die Messinvarianzprüfung für die Sicherheitsskala zeigt, dass die Skala mit allen Items lediglich die metrische Messinvarianz auf akzeptablen Niveau unterstützt. Zwar lässt sich durch die Aufhebung der Beschränkungen für die Items 2 und 5 der CFI steigern. Jedoch zeigt auch das partial-skalare Messinvarianzmodell keinen akzeptablen CFI (= 0.853) (Tab. 5.20).

Hinsichtlich der MIMIC-Prüfung erweisen sich die Daten der Items 2 und 6 im UDA-Assessment heterogen in der Stichprobe. Der latente Faktor erweist sich als messinvariant (Tab. 5.21).

Rechtfertigungsskala

Auch die NOS-Skala Rechtfertigung erfüllt nur das weniger konservative Evaluationskriterium (CFI = 0.929). Die Daten der *Rechtfertigungsskala* unterstützen jedoch selbst

Tabelle 5.21: Regressionen im MIMIC-Ansatz zur Sicherheitsskala. Dargestellt sind die Regressionen der Lernumgebung auf den latenten Faktor und die manifesten Indikatoren in einem konfiguralen Messinvarianzmodell.

	UDA-Assessment		Originalssessment	
	Schätzung	p-Wert	Schätzung	p-Wert
Latenter Faktor	0.20	0.10	0.02	0.89
Item 1	0.20	0.21	-0.01	0.96
Item 2	0.39	0.01	0.06	0.69
Item 3	0.23	0.10	0.07	0.63
Item 4	-0.21	0.18	0.04	0.79
Item 5	-0.02	0.87	0.05	0.74
Item 6	-0.41	0.00	0.23	0.08
Item 7	0.18	0.23	-0.16	0.28

Anmerkung:
CFI: 0.902; TLI: 0.861; RMSEA: 0.00; SRMR: 0.063

partial-strikte Messinvarianz (CFI = 0.901). Die Ausnahme bildet das Item 3, was im Besonderen durch die Lernumgebungen der Intervention adressiert wurde (Tab. 5.22).

Letzlich erweist sich auch bei der *Rechtfertigungsskala* der latente Faktor als messinvariant und wird nicht signifikant durch die Lernumgebung regressiert (Tab. 5.23).

Entwicklungsskala

Für die gesamte Entwicklungsskala zeichnet sich ein vergleichbares Bild. Die Fitindizes verletzen in den ersten Messinvarianzprüfungen nur knapp nicht das weniger konservative Qualitätskriterium für den CFI. Die Daten unterstützen jedoch in diesem Fall keine partial-skalare MI (Tab. 5.24). Allerdings zeigen sich für die Art der Lernumgebung auch keine signifikanten Regressionen bezüglich eines der Items.

Insgesamt erweisen sich die Vollskalen als nicht geeignet, um änderungen in den NOS-Konzepten für die einzelnen Skalen abzubilden. Dies kann an der kleinen Stichprobe liegen, jedoch zeigen bereits die Voranalysen zur internen Konsistenz der Skalen beider Assessmentversionen, dass diese teilweise nicht ausreichend ist. Die unzureichende interne Konsistenz wirkt sich nun auf die longitudinale MI-Prüfung der Vollskalen aus.

5.2.2.3 Bestimmung der standardisierten Faktorladungen und Berechnung von T-Tests für jedes Item im Prä-Post-Vergleich

Für die Analyse ist es von Bedeutung festzulegen, welche der Items der NOS-Skalen tatsächlich mit den Lernumgebungen verbunden sind. Da für jede Skala zumindest longitudinale metrische MI vorlag, werden im Folgenden T-Tests für verbundene Stichproben durchgeführt, um Mittelwertsunterschiede der Skalen zu bestimmen (Tab. 5.26).

Außerdem werden die standardisierten Faktorladungen bestimmt (Tab. 5.27). Aus der Kombination beider Analysen ergibt sich, dass gerade die Items mit einer niedrigen Faktorladung auch diejenigen sind, die in keiner der beiden Assessmentformen eine signifikante

Tabelle 5.22: Messinvarianzmodelle zur NOS-Skala Rechtfertigung

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	206.6	138	<0.05	0.065	0.929	0.906	0.061	Ja
Metrisch	227.48	156	<0.05	0.062	0.926	0.914	0.077	Ja
Skalar	274.02	174	<0.05	0.07	0.896	0.892	0.083	Nein
Partial skalar	259.89	172	<0.05	0.066	0.909	0.904	0.083	Ja
Strikt	285.33	190	<0.05	0.065	0.901	0.905	0.086	Ja

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für die Mittelwerte von Item 3 aufgehoben
Angen.?: Angenommen?

Mittelwertsänderung zeigen. Für die Neubildung der Skalen werden einerseits Items mit einer ausreichend hohen standardisierten Faktorladung ausgewählt. Andererseits wurden auch Items mit signifikanten Mittelwertsänderungen trotz einer geringen standardisierten Faktorladung mitaufgenommen (Tab. 5.27).

Tabelle 5.23: Regressionen im MIMIC-Ansatz zur Rechtfertigungsskala. Dargestellt sind die Regressionen der Lernumgebung auf den latenten Faktor und die manifesten Indikatoren in einem konfiguralen Messinvarianzmodell.

	UDA-Assessment		Originalssessment	
	Schätzung	p-Wert	Schätzung	p-Wert
Latenter Faktor	-0.16	0.09	-0.05	0.56
Item 1	-0.21	0.21	0.19	0.30
Item 2	0.00	0.97	0.11	0.48
Item 3	0.00	0.97	0.07	0.63
Item 4	0.03	0.82	0.10	0.40
Item 5	0.18	0.19	-0.05	0.66
Item 6	0.04	0.77	-0.20	0.07
Item 7	-0.14	0.26	-0.12	0.34

Anmerkung:
CFI: 0.924; TLI: 0.892; RMSEA: 0.00; SRMR: 0.062

Tabelle 5.24: Messinvarianzmodelle zur NOS-Skala Entwicklung

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	343.39	190	<0.05	0.085	0.903	0.878	0.065	Ja
Metrisch	360.14	211	<0.05	0.08	0.906	0.893	0.072	Ja
Skalar	452.79	232	<0.05	0.093	0.861	0.856	0.081	Nein
Partial skalar	391.18	223	<0.05	0.082	0.894	0.886	0.077	Nein
Strikt	508.56	241	<0.05	0.1	0.831	0.832	0.09	Nein

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für die Faktorladungen für die
Items 1und 6 zu beiden Messzeitpunkten 1 und 2.
Angen.?: Angenommen?

Tabelle 5.25: Regressionen im MIMIC-Ansatz zur Entwicklungsskala. Dargestellt sind die Regressionen der Lernumgebung auf den latenten Faktor und die manifesten Indikatoren in einem konfiguralen Messinvarianzmodell.

	UDA-Assessment		Originalssessment	
	Schätzung	p-Wert	Schätzung	p-Wert
Latenter Faktor	0.11	0.34	-0.09	0.45
Item 1	-0.03	0.86	-0.19	0.13
Item 2	-0.05	0.73	0.06	0.65
Item 3	0.09	0.57	0.03	0.75
Item 4	-0.08	0.55	0.06	0.59
Item 5	-0.11	0.39	-0.02	0.90
Item 6	0.01	0.97	-0.03	0.82
Item 7	0.22	0.17	0.01	0.91
Item 8	0.11	0.52	0.08	0.45

Anmerkung:
CFI: 0.894; TLI: 0.859; RMSEA: 0.051; SRMR: 0.068

Tabelle 5.26: Signifikante Mittelwertsänderung der Items in beiden Assessments (Bonferroni-Korrektur durchgeführt).

	UDA-Assessment		Originalssessment	
Item	MZP 2 - MZP 1	p	MZP 2 - MZP 1	p
Herkunftsskala				
Item 1	-0.19	1.000	0.06	1.000
Item 2	-0.35	0.108	-0.04	1.000
Item 3	0.61	<0.05	0.35	<0.05
Item 4	-0.17	1.000	-0.24	0.378
Item 5	-0.54	<0.05	-0.03	1.000
Sicherheitsskala				
Item 1	-0.11	1.000	0.07	1.000
Item 2	-0.38	<0.05	-0.11	1.000
Item 3	-0.11	1.000	-0.01	1.000
Item 4	-0.39	<0.05	-0.10	1.000
Item 5	-0.34	<0.05	-0.16	1.000
Item 6	-0.75	<0.05	-0.06	1.000
Item 7	-0.49	<0.05	-0.16	1.000
Entwicklungsskala				
Item 1	0.49	<0.05	0.16	1.000
Item 2	0.47	<0.05	0.26	0.135
Item 3	0.46	<0.05	-0.11	1.000
Item 4	0.11	1.000	0.08	1.000
Item 5	0.26	0.243	0.19	0.729
Item 6	0.27	0.189	0.07	1.000
Item 7	-0.21	1.000	-0.22	0.513
Item 8	-0.12	1.000	-0.12	1.000
Rechtfertigungsskala				
Item 1	0.45	<0.05	-0.02	1.000
Item 2	0.12	1.000	-0.24	0.378
Item 3	0.17	1.000	0.00	1.000
Item 4	0.44	<0.05	0.18	1.000
Item 5	0.22	0.297	0.04	1.000
Item 6	-0.09	1.000	0.01	1.000
Item 7	0.17	1.000	0.10	1.000

Anmerkung:
MZP: Messzeitpunkt

Tabelle 5.27: Neuformulierte NOS-Skalen mit den stand. Faktorladungen, Mittelwertsdifferenzen und den dazugehörigen, Bonferroni-korrigierten Signifikanzen.

	UDA-Assessment				Originalassessment			
	stand. Faktorladung		Mittelwerte		stand. Faktorladung		Mittelwerte	
	MZP 1	MZP 2	MZP 2 - MZP 1	p-Wert	MZP 1	MZP 2	MZP 2 - MZP 1	p-Wert
Herkunftsskala								
Item 2	0.27	0.77	-0.35	0.00	0.50	0.66	-0.04	0.71
Item 3	0.54	0.65	0.61	0.00	0.37	0.71	0.35	0.00
Item 4	0.60	0.79	-0.17	0.16	0.78	0.78	-0.24	0.01
Item 5	0.52	0.69	-0.54	0.00	0.69	0.63	-0.03	0.74
Sicherheitsskala								
Item 3	0.65	0.75	-0.11	0.29	0.66	0.70	-0.01	0.90
Item 4	0.82	0.73	-0.39	0.00	0.55	0.70	-0.10	0.33
Item 6	0.64	0.71	-0.75	0.00	0.63	0.63	-0.06	0.50
Item 7	0.62	0.73	-0.49	0.00	0.66	0.62	-0.16	0.10
Entwicklungsskala								
Item 1	0.62	0.72	0.49	0.00	0.56	0.66	0.16	0.10
Item 2	0.66	0.73	0.47	0.00	0.64	0.71	0.26	0.00
Item 4	0.66	0.80	0.11	0.18	0.65	0.76	0.08	0.41
Item 5	0.73	0.78	0.26	0.01	0.67	0.65	0.19	0.03
Rechtfertigungsskala								
Item 2	0.72	0.70	0.12	0.26	0.25	0.38	-0.24	0.01
Item 3	0.52	0.60	0.17	0.08	0.21	0.46	0.00	1.00
Item 4	0.63	0.69	0.44	0.00	0.44	0.71	0.18	0.06
Item 6	0.66	0.68	-0.09	0.36	0.50	0.73	0.01	0.88

Anmerkung:
MZP: Messzeitpunkt

Tabelle 5.28: Interne Konsistenzen der Kurzskalen: McDonalds Omega

NOS-Skala	Messzeitpunkt 1	Messzeitpunkt 2	Beide Messzeitpunkte
UDA-Assessment			
Herkunft	0.86	0.89	0.9
Sicherheit	0.8	0.81	0.85
Entwicklung	0.78	0.85	0.87
Rechtfertigung	0.74	0.78	0.8
Originalassessment			
Herkunft	0.83	0.9	0.89
Sicherheit	0.74	0.75	0.81
Entwicklung	0.75	0.79	0.82
Rechtfertigung	0.41	0.66	0.62

Die erneute Analyse zur internen Konsistenz zeigt, dass bis auf die *Rechtfertigungsskala* im Originalassessment, die Werte weiterhin im akzeptablen bis guten Bereich liegen (Tab. 5.28). Auch steigt die Konsistenz erneut vom ersten zum zweiten Messzeitpunkt. Durch die Verkürzung der Skalen wird eine erneute longitudinale Messinvarianzprüfung notwendig.

5.2.2.4 Messinvarianzprüfung der einzelnen Kurzskalen

Herkunftsskala

Für die *verkürzte Herkunftsskala* lassen sich konfigurale und metrische MI feststellen (Tab. 5.29). Dies wird auch durch den χ^2-Differenztest bestätigt ($\Delta\chi^2 = 12.61$, $p =$ n.s.). Die restriktivere skalare MI unterstützen die Daten nicht (ΔCFI $= 0.014$ und $\Delta\chi^2 = 43.86$, $p < .000$). Wie bereits auf manifester Ebene aus Tab. 5.26 ersichtlich ist, ist Item 3 der Skala das einzige, was sich in beiden Assessmentversionen dieser Skala signifikant ändert. Entsprechend weist auch die Kurzskala eine Messvarianz für dieses Item auf. Durch die Aufhebung der Beschränkung der Faktorladung zum ersten Messzeitpunkt für das Item 3 der Herkunftsskala unterstützen die Daten partial-skalare MI (CFI $= 0.955$). Strikte MI wird von den Daten auf Basis der CFI-Differenzen nicht unterstützt jedoch über den χ^2-Differenztest (ΔCFI $= 0.111$ und $\Delta\chi^2 = 20.11$, $p =$ n.s.). Mit Blick auf die große Differenz zwischen dem partial-skalaren Messmodell und dem strikten Messmodell wird aber keine strikte MI angenommen.

Sicherheitsskala

Auch die *verkürzte Sicherheitsskala* unterstützt die konfigurale MI. Das CFI-Kriterium wird allerdings für die metrische MI knapp verletzt (ΔCFI $= 0.012$). Allerdings weisen die anderen Fitindizes gute Werte auf (Tab. 5.30). Außerdem unterstützt der χ^2-Differenztest die Gleichheit der Datenstrukturen ($\Delta\chi^2 = 13.71$, $p =$ n.s.). Entsprechend wird davon ausgegangen, dass metrische MI vorliegt. Die skalare MI wird allerdings von keinem der Indizes oder Tests voll unterstützt (ΔCFI $= 0.065$; $\Delta\chi^2 = 33.08$, $p < .000$). Durch die Aufhebung der Faktorladungen von Item 1 und 6 zum ersten Messzeitpunkt wird partial-skalare (CFI $= 0.945$) und auch strikte MI von den Daten unterstützt (ΔCFI $= 0.003$; $\Delta\chi^2 = 6.87$, $p =$ n.s.).

Tabelle 5.29: Messinvarianzmodelle zur verkürzten NOS-Kurzskala Herkunft.

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	37.35	30	0.167	0.045	0.985	0.972	0.041	Ja
Metrisch	53.18	39	0.065	0.054	0.971	0.958	0.068	Ja
Skalar	114.8	48	<0.05	0.106	0.863	0.84	0.094	Nein
Partial-skalar	64.15	46	<0.05	0.057	0.963	0.955	0.074	Ja
Strikt	136.96	58	<0.05	0.105	0.838	0.844	0.096	Ja

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für Faktorladung von Item 3 wurde zum ersten Messzeitpunkt aufgehoben; Angen.?: Angenommen?

Tabelle 5.30: Messinvarianzmodelle zur verkürzten NOS-Kurzskala Sicherheit.

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	55.46	30	<0.05	0.083	0.957	0.92	0.049	Ja
Metrisch	71.53	39	<0.05	0.082	0.945	0.921	0.068	Ja
Skalar	119.1	48	<0.05	0.109	0.88	0.86	0.098	Nein
Partial-skalar	76.69	44	<0.05	0.077	0.945	0.93	0.071	Ja
Strikt	87.18	56	<0.05	0.067	0.948	0.948	0.067	Ja

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für Faktorladung von Item 3 und 6 wurde zum ersten Messzeitpunkt aufgehoben; Angen.?: Angenommen?

Entwicklungsskala

Die *verkürzte Entwicklungsskala* zeigt bessere Fitindizes (Tab. 5.31). Die Daten unterstützen sowohl konfigurale und metrische MI. Auch hier unterstützen die Daten jedoch nicht die volle skalare MI (ΔCFI = 0.017; $\Delta\chi^2 = 20.4$, $p < .05$). Die Beschränkung der Faktorladung von Item 2 zum ersten Messzeitpunkt sorgt für die Messvarianz. Mit ihrer Aufhebung kann partial-skalare MI angenommen werden (CFI = 0.971). Strikte MI wird von den Daten nicht unterstützt (ΔCFI = 0.023; $\Delta\chi^2 = 39.72$, $p < .000$).

Rechtfertigungsskala

Die *verkürzte Rechtfertigungsskala* zeigt ebenfalls bessere Fitindizes (Tab. 5.32). Die Daten unterstützen sowohl konfigurale, metrische (ΔCFI = -0.007; $\Delta\chi^2 = 5.87$, p = n.s.) als auch die volle skalare MI (ΔCFI = 0.002; $\Delta\chi^2 = 8.55$, p = n.s.). Für die strikte MI wird das Qualitätskriterium verletzt (ΔCFI = 0.013), jedoch stellt der χ^2-Differenztest keinen signfikanten Unterschied für die Modell- und Datenstrukturen fest ($\Delta\chi^2 = 12.73$,

Tabelle 5.31: Messinvarianzmodelle zur verkürzten NOS-Kurzskala Entwicklung.

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	41.31	30	0.082	0.053	0.984	0.971	0.034	Ja
Metrisch	48.93	39	0.132	0.043	0.986	0.98	0.046	Ja
Skalar	69.95	48	<0.05	0.058	0.969	0.964	0.057	Nein
Partial-Skalar	63.19	46	<0.05	0.052	0.976	0.971	0.054	Ja
Strikt	87.18	56	<0.05	0.067	0.948	0.948	0.067	Nein

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Partial-skalar: Beschränkung für Faktorladung von Item 2 zum ersten Messzeitpunkt
Angen.?: Angenommen?

Tabelle 5.32: Messinvarianzmodelle zur verkürzten NOS-Kurzskala Rechtfertigung.

				Fit-Werte				
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR	Angen.?
Konfigural	41.31	30	0.082	0.053	0.984	0.971	0.034	Ja
Metrisch	48.93	39	0.132	0.043	0.986	0.98	0.046	Ja
Skalar	69.95	48	<0.05	0.058	0.969	0.964	0.057	Ja
Strikt	70.96	59	0.137	0.039	0.976	0.977	0.073	Ja

Note:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Angen.?: Angenommen?

p = n.s.). Und weiterhin liegen die übrigen Fitindizes im guten bis sehr guten Bereich. Entsprechend wird angenommen, dass die Daten strikte MI unterstützen.

5.2.3 Prüfung der Testzugänglichkeit der Assessmentversionen

Zur Prüfung der Testzugänglichkeit wurden die Daten aus beiden gekürzten NOS-Assessments zum ersten Messzeitpunkt verwendet (Tab. 5.28). Für die Analyse wird das R-Paket `pairwise` verwendet (Heine, 2017). Zum zweiten Messzeitpunkt kann bereits eine Beeinflussung durch die Lernumgebung vorliegen. Um die Gruppenabhängigkeit der Items zu untersuchen, wurden die Lernendenmerkmale *Lesefähigkeit*, *Intelligenz*, *Geschlecht* und *sozioökonomischer Status* genutzt. Alle Lernendenmerkmale wurden unabhängig vom Studiendesign mit den gleichen Instrumenten erhoben. Bei keiner der Variablen liegt eine Normalverteilung vor (*Lesefähigkeit*: W = 0.96, $p < .000$; *Intelligenz*: W = 0.93, $p < .000$; *sozioökonomischer Status*: W = 0.75, $p < .000$;). Während der *sozioökonomische Status* vergleichbar zu anderen europäischen Staaten ist (Torsheim et

Tabelle 5.33: Statistisches Beschreibungen für die Lernendenmerkmale

	M	SD	Max.	Min.	Cut-off Wert	n < Cut-off	n > Cut-off
UDA-Assessment							
Lesefähigkeit	20.98	5.11	15.87	36	7	16	159
Intelligenz	13.97	7.38	6.59	25	1	37	138
SES	8.77	2.62	6.15	12	0	21	154
Originalassessment							
Lesefähigkeit	20.16	4.03	16.14	34	12	33	132
Intelligenz	12.90	7.24	5.66	25	1	36	129
SES	8.71	2.82	5.88	12	0	16	149

Anmerkung:
Als Kriterium für die Cut-off-Werte wurde eine Standardabweichungen gewählt.
UDA-Assessment: Weiblich = 79, Männlich = 96
Originalsassessment: Weiblich = 93, Männlich = 72
SES: Sozioökonomischer Status
Max. : Maximum; Min.: Minimum

al., 2016), folgt die Verteilung der KFT-Rohwerte am ehesten der Normierungsstichprobe für die Hauptschule (Heller & Perleth, 2000). Die Rohwerte für die Lesefähigkeit sind im Vergleich zur Normierungsstichprobe zu niedrig (Wimmer & Mayringer, 2014). Eine übertragung der Rohwerte zu den T-Werten würde zu ca. 25 % bzw. 30 % fehlenden Werten führen. Aus diesen Gründen werden die Analysen mit den Rohwerten und nicht mit den T-Werten der Testmanuale durchgeführt. Allerdings ist damit auch kein Vergleich zur Normierungsstichprobe möglich.

In einem ersten Schritt wurde der Mittelwert für die *Lesefähigkeit*, die *Intelligenz* und den *sozioökonomischen Status* gebildet (Tab. 5.33). Der Anteil von Kindern mit *sonderpädagogischem Unterstützungsbedarf* ist zu gering, als dass hier eine eigene Analyse vorgenommen werden könnte. Jedoch fallen viele Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf in die folgende Definition von Risikolernenden; besonders häufig betrifft dies die Lesefähigkeit. Zur Gruppenbildung der Risikolernenden wurde eine Differenz von einer Standardabweichung ausgehend vom Mittelwert als Cut-of-Kriterium gewählt. Eine größere Abweichung ist aufgrund der sehr niedrigen Mittelwerte für die *Intelligenz* und die *Lesefähigkeit* nicht möglich. Bezüglich der statistischen Kennwerte lassen sich keine signifikanten Unterschiede für beide Assessmentversionen beobachten (*Lesefähigkeit*: $t(328) = 1.65$, p = n.s.; *Intelligenz*: $t(337) = 1.34$, p = n.s.; *Sozioökonomischer Status*: $t(332) = 0.21$, p = n.s.). über diese Kriterien können insgesamt 137 Risikolernende identifiziert werden, die mindestens ein Kriterium erfüllen. Dies entspricht 40 % der Gesamtstichprobe.

Testzugänglichkeit der Herkunftsskala

Für das *UDA-Assessment* lassen sich hinsichtlich der vier Gruppenvergleiche alle Items der *Herkunftsskala* als messinvariant beschreiben (Abb. 5.1). Auffällig ist das Item drei, was auf der latenten Metrik schwieriger als die Items zwei, vier und fünf ist. Letztere weisen einen in etwa gleichen Schwierigkeitsgrad auf (Formulierung von Item 3: "Nur Naturwissenschaftler können naturwissenschaftliche Theorien entwickeln."). Für das *Originalassessment* erweisen sich fast alle Items in den Gruppenvergleichen zum Geschlecht,

der Intelligenz und der Lesefähigkeit als messinvariant (Abb. 5.2). Im Vergleich zum sozioökonomischen Status gilt dies für Item 3 nicht. Dieses ist für Schülerinnen und Schüler mit geringem sozioökonomischen Status schwieriger zu beantworten.

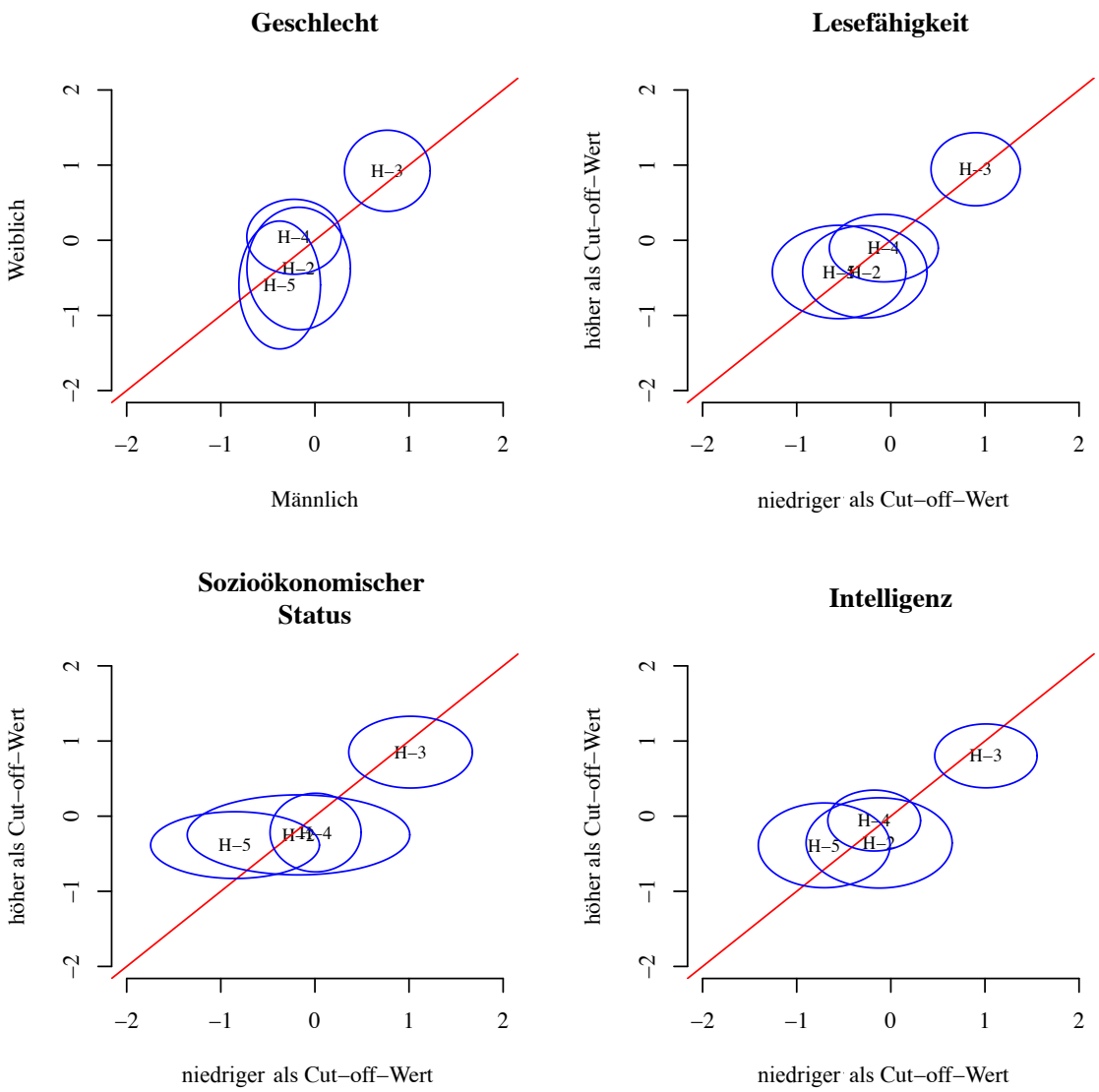


Abbildung 5.1: DIF-Analyse der Herkunftsskala aus dem UDA-Assessment.

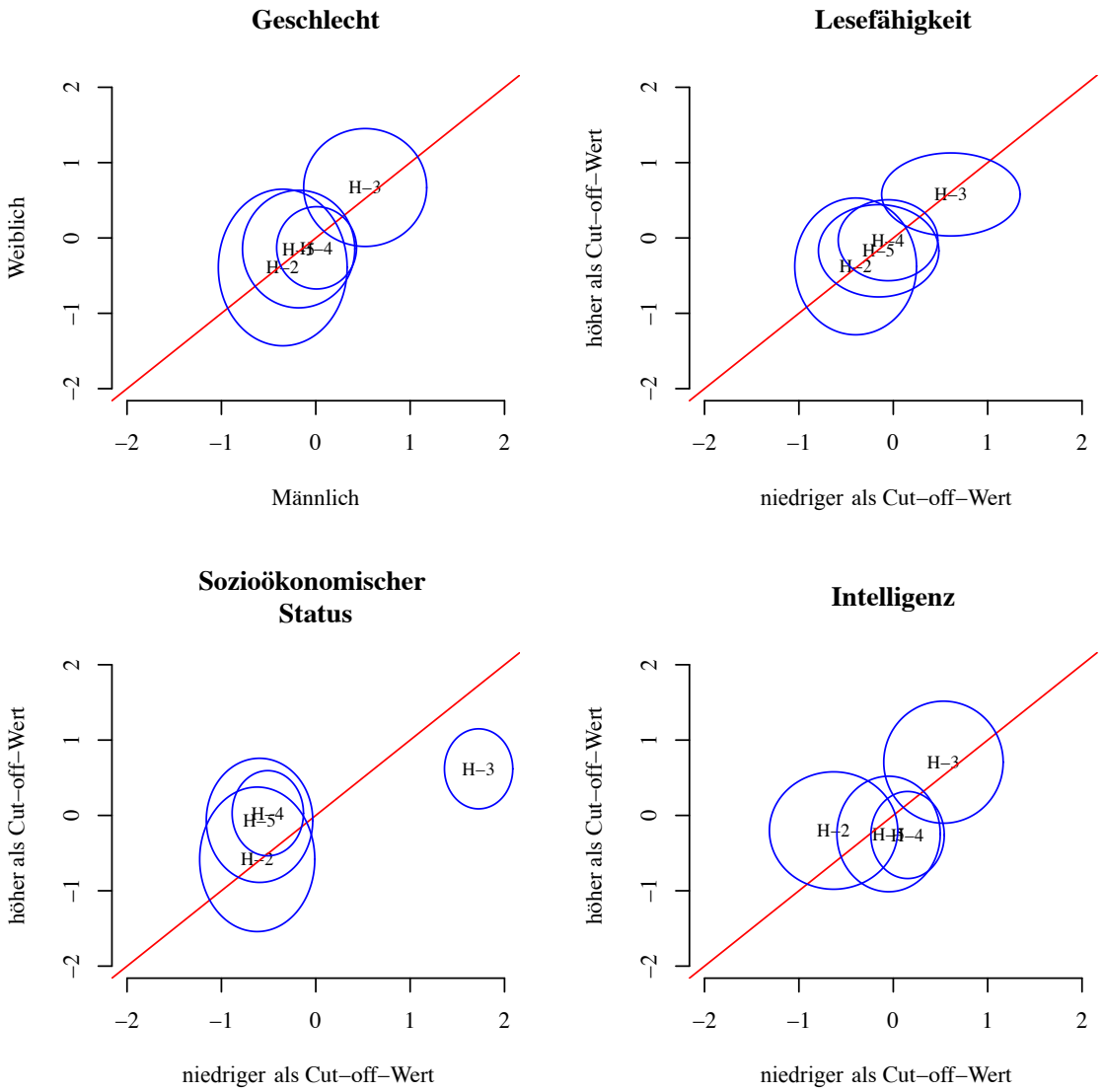


Abbildung 5.2: DIF-Analyse der Herkunftsskala aus dem Originalassessment.

Testzugänglichkeit der Sicherheitsskala

Auch die Items der NOS-Skala *Sicherheit* erweisen sich im UDA-Assessment in allen vier Gruppenvergleichen als messinvariant (Abb. 5.3). Auffällig ist das Item vier (“Für alle Fragen in den Naturwissenschaften gibt es immer nur eine Lösung.”), was leichter korrekt beantwortet wird als die anderen Items (Item drei, vier und sieben). Im Originalassessment zeigt nur das Item 6 eine Gruppenabhängigkeit zum sozioökonomischen Status (“Naturwissenschaftler stimmen immer darin überein, was in ihrem Fach wahr ist.”) (Abb. 5.4). Schülerinnen und Schüler mit höherem *sozioökonomischen Status* lehnen diese Aussagen eher ab. Die Itemschwierigkeiten untereinander sind in jedem Gruppenvergleich vergleichbar.

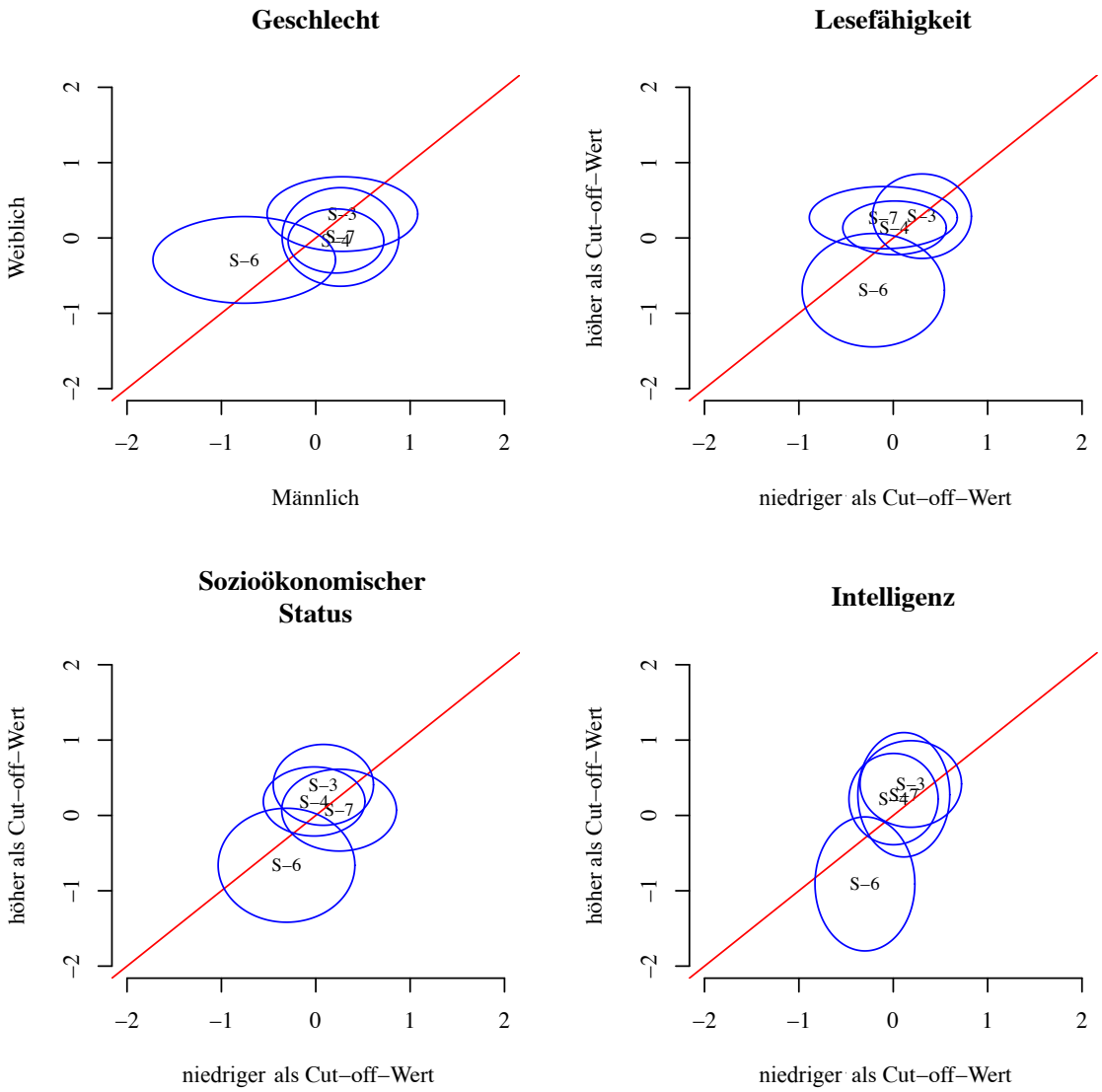


Abbildung 5.3: DIF-Analyse der Sicherheitsskala aus dem UDA-Assessment.

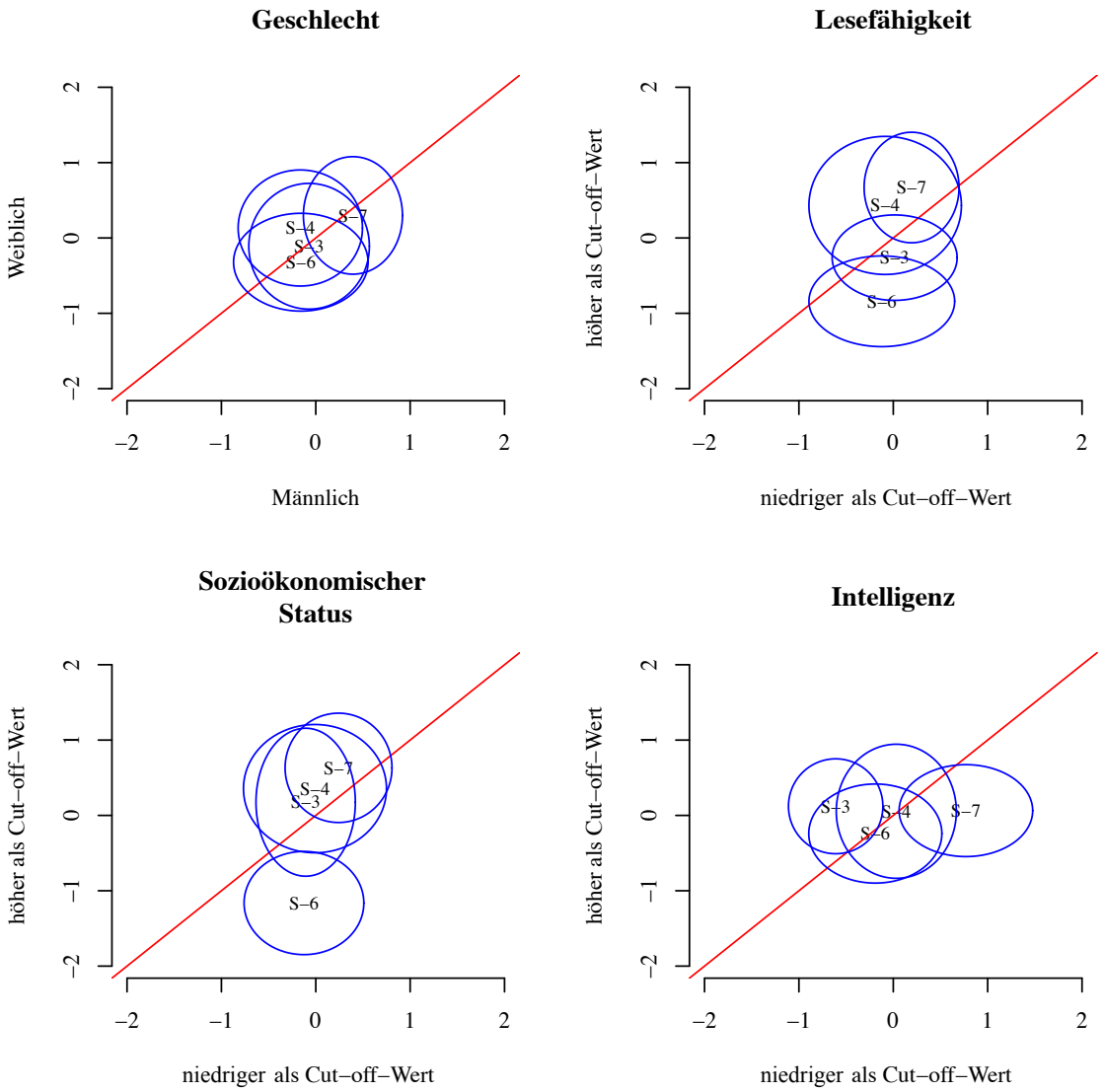


Abbildung 5.4: DIF-Analyse der Sicherheitsskala aus dem Originalassessment.

Testzugänglichkeit der Entwicklungsskala

Auch die Items des UDA-Assessments der *Entwicklungsskala* zeigen Messinvarianz in allen vier Gruppenvergleichen (Abb. 5.5). Die Itemschwierigkeiten sind vergleichbar, wobei das Item eins ("Naturwissenschaftler ändern manchmal ihre Meinung darüber, was in ihrem Fach wahr ist.") etwas schwieriger und das Item 4 etwas leichter ist ("Naturwissenschaftler verändern oder ersetzen naturwissenschaftliche Theorien, wenn neue Beweise vorliegen."). Im Originalassessment erweisen sich nur die Items 2 und 4 in allen Gruppenvergleichen als messinvariant (Abb. 5.6). Item 2 ist dabei durchgehend leichter als die übrigen Items dieser Skala. Item 5 ist gruppenabhängig hinsichtlich des Geschlechts und der Lesefähigkeit ("Manchmal verändern sich die Vorstellungen in den Naturwissenschaften."). Es wird eher von Mädchen richtig gelöst und eher weniger oft von Schülerinnen und Schülern mit einer niedrigen Lesefähigkeit. Dagegen wird Item 1 ("Manchmal ändern Naturwissenschaftler ihre Meinung darüber, was in ihrem Fach wahr ist.") eher von Jungs richtig beantwortet.

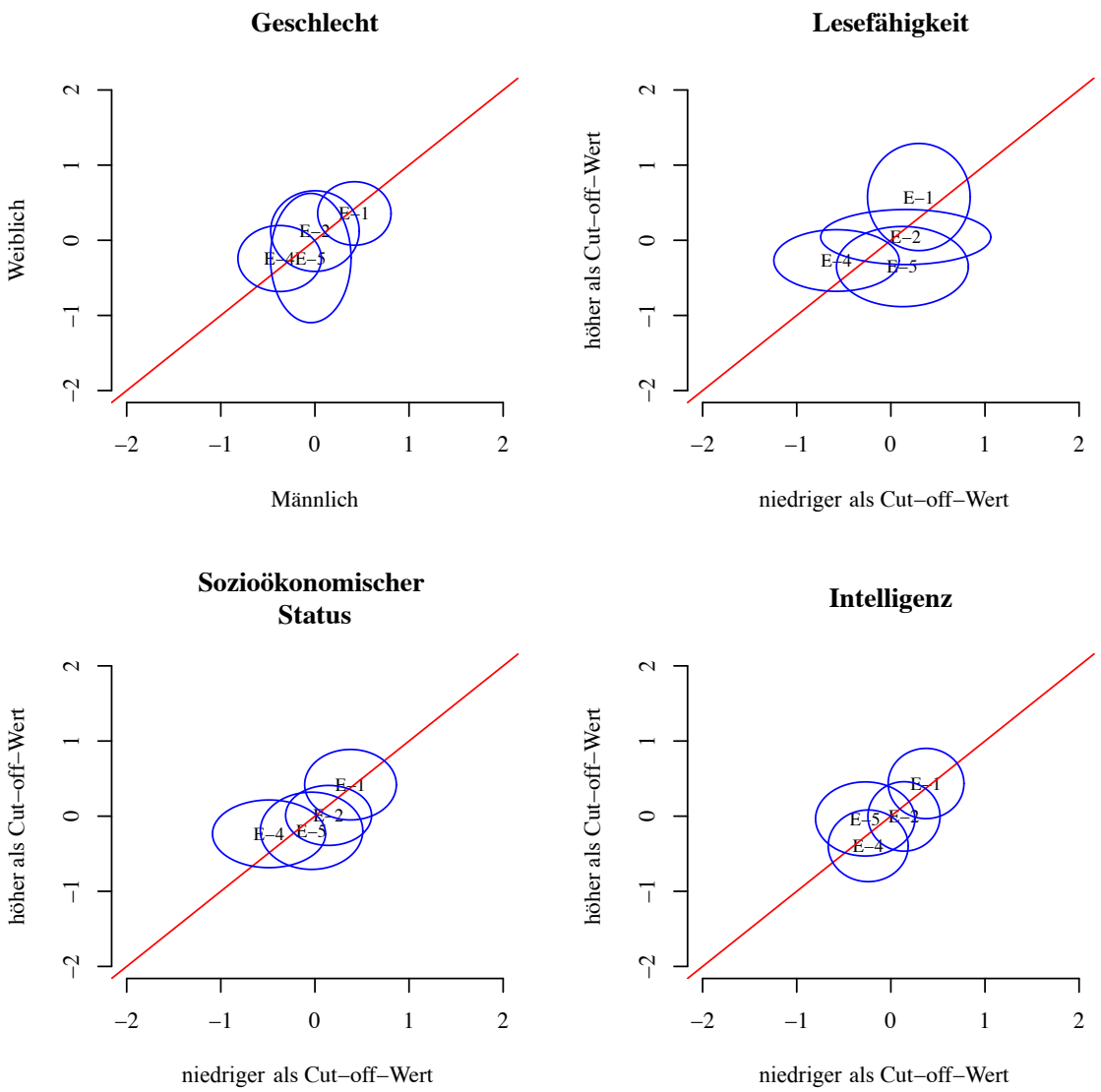


Abbildung 5.5: DIF-Analyse der Entwicklungsskala aus dem UDA-Assessment.

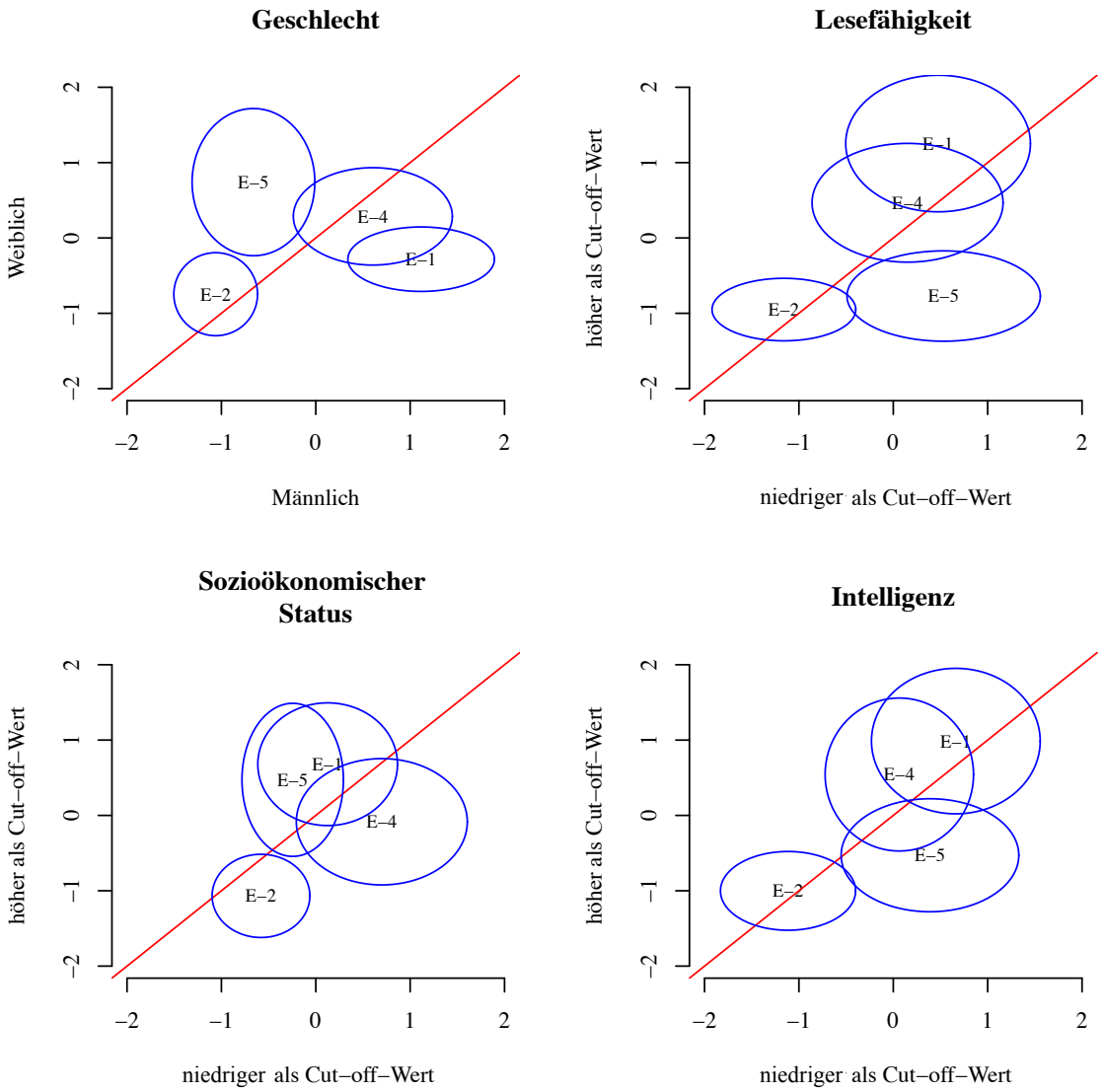


Abbildung 5.6: DIF-Analyse der Entwicklungsskala aus dem Originalassessment.

Testzugänglichkeit der Rechtfertigungsskala

Schließlich zeigen auch die Items der *Rechtfertigungsskala* des UDA-Assessments Messinvarianz in allen vier Gruppenvergleichen (Abb. 5.7). Die Itemschwierigkeiten sind auch hier vergleichbar. Im Originalassessment zeigt sich mit Ausnahme des Items zwei ein ähnliches Bild (Abb. 5.8). Das Item 2 (“Wenn Naturwissenschaftler Experimente durchführen, legen sie im Voraus einige Aspekte der Untersuchung fest.”) lösen Schülerinnen und Schüler mit einer höheren Intelligenz eher als mit einer niedrigen. Außerdem führt ein niedriger sozioökonomischer Status zu einer höheren Lösungswahrscheinlichkeit.

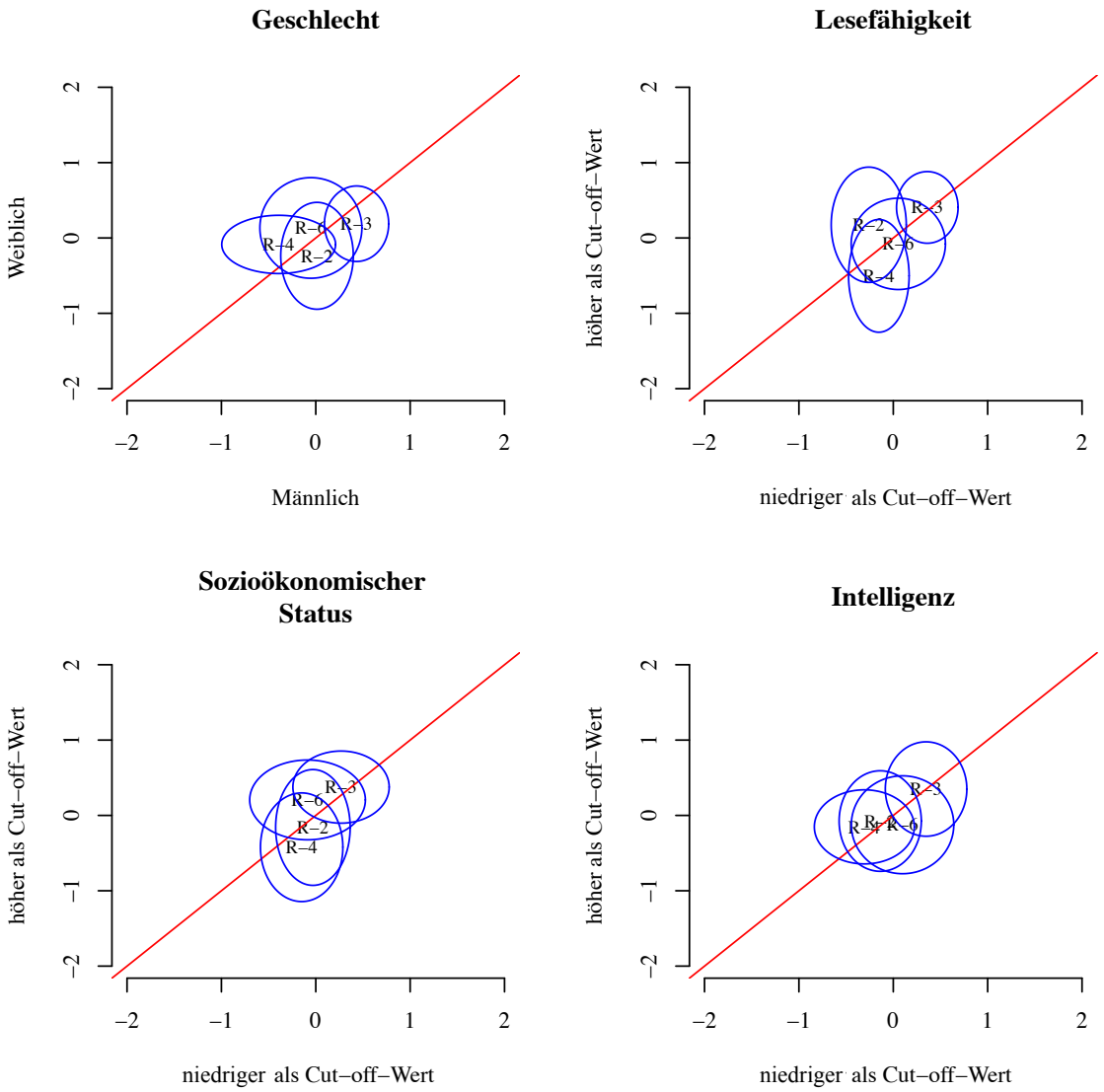


Abbildung 5.7: DIF-Analyse der Rechtfertigungsskala aus dem UDA-Assessment.

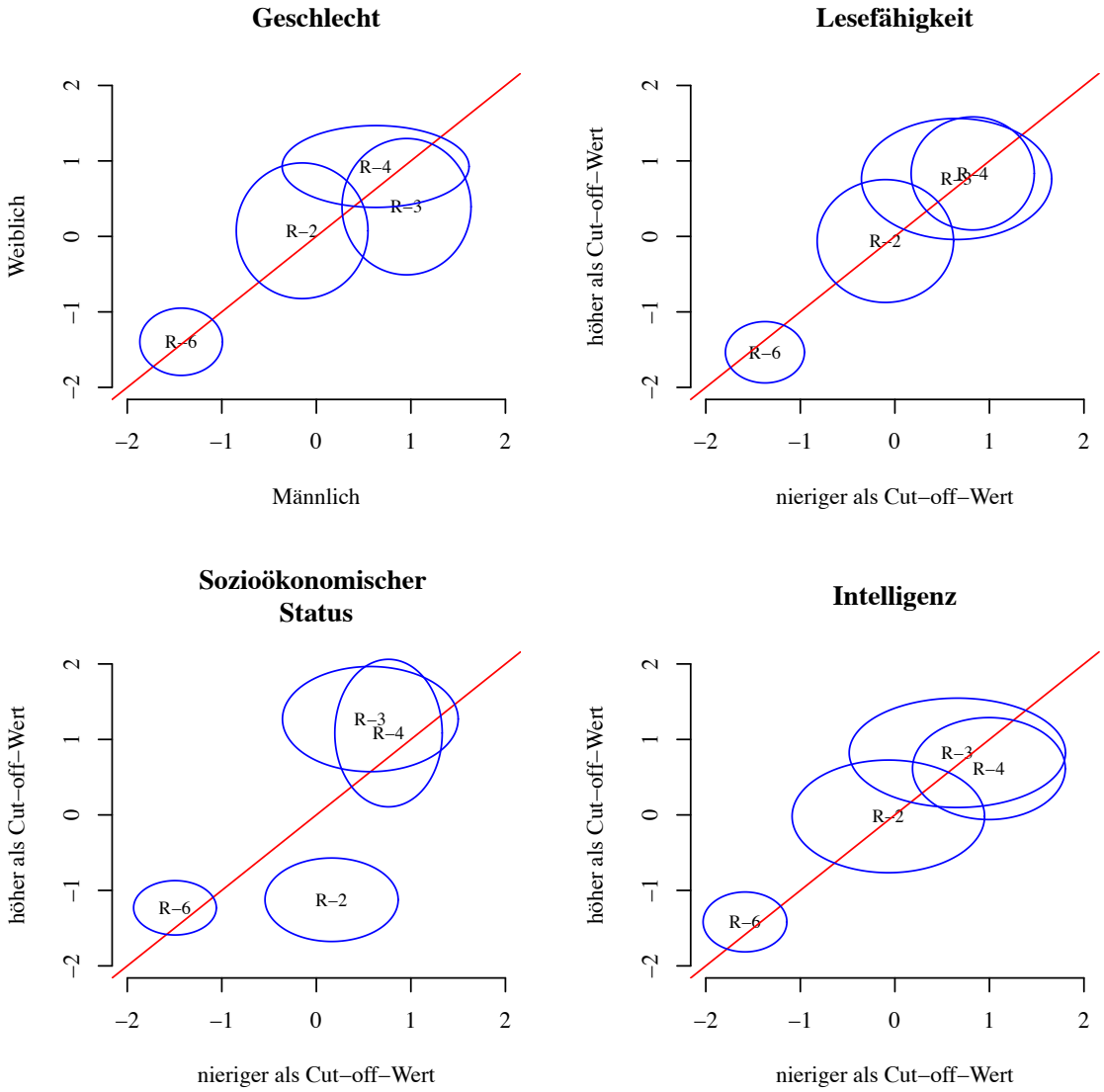


Abbildung 5.8: DIF-Analyse der Rechtfertigungsskala aus dem Originalassessment.

Insgesamt erweist sich das UDA-Assessment zum ersten Messzeitpunkt als testzugänglicher als das Originalassessment. Allerdings sind die Unterschiede nicht gravierend, was bei einer Likert-Skala auch nicht unbedingt zu erwarten ist.

5.2.4 Untersuchung auf einen möglichen Differential Boost

Zur Untersuchung auf einen *Differential Boost* werden einerseits Methoden zur Bestimmung von Mittelwertsunterschieden in Kombination mit Effektstärkenbestimmungen vorgeschlagen (Phillips, 1994). Elliott et al. (2018) schlagen andererseits aber auch die Bestimmung von internen Konsistenzen vor, um weitergehende Auswirkungen von Testadaptionen zu bestimmen. Demnach sollte die interne Konsistenz für Risikogruppen nicht geringer sein als die Werte der Vergleichsgruppe. Ist die interne Konsistenz größer als bei der Vergleichsgruppe liegt ein *Differential Boost* vor. Als *Risikolernende* gelten in dieser Studie Schülerinnen und Schüler, die mindestens einmal zu einer Risikogruppen gehören (Tab. 5.33). Die Fit-Werte dieser Modelle sind trotz der kleineren Gruppe im guten bis

sehr guten Bereich (Tab. 5.35). Eine Ausnahme bildet die *Entwicklungsskala* im *Originalassessment*. Hier sind die CFI und die TLI-Werte mäßig bis akzeptabel. Allerdings sind der RMSEA und der SRMR-Wert im guten bis sehr guten Bereich. Für die Gruppenvergleiche kann daher festgehalten werden, dass grundsätzlich die gleichen Konstrukte innerhalb der Substichproben zum ersten Messzeitpunkt gemessen werden.

Tabelle 5.34: Interne Konsistenzen (McDonalds-ω) der Kurzskalen getrennt nach Risikolernenden und Anderen zum ersten Messzeitpunkt

	UDA-Assessment		Originalassessment	
NOS-Skala	Risikolernende	Andere	Risikolernende	Andere
Herkunft	0.53	0.6	0.62	0.69
Sicherheit	0.72	0.37	0.35	0.4
Entwicklung	0.85	0.76	0.73	0.69
Rechtfertigung	0.83	0.67	0.55	0.4

Tabelle 5.35: Konfigurale Messinvarianzmodelle je Assessment im Vergleich zwischen Riskolernenden und Anderen zum ersten Messzeitpunkt.

				Fit-Werte			
Stufe	Chi-Quadrat	dF	p-Wert	RMSEA	CFI	TLI	SRMR
UDA-Assessment							
Herkunft	3.37	4	0.498	0	1	1.037	0.028
Sicherheit	1.04	4	0.903	0	1	1.052	0.012
Entwicklung	2.95	4	0.566	0	1	1.016	0.017
Rechtfertigung	4.31	4	0.365	0.03	0.998	0.994	0.025
Originalssessment							
Herkunft	9.5	4	0.050	0.129	0.945	0.834	0.036
Sicherheit	4.3	4	0.367	0.03	0.998	0.993	0.026
Entwicklung	10.93	4	<0.05	0.145	0.943	0.828	0.04
Rechtfertigung	1.22	4	0.874	0	1	1.411	0.018

Anmerkung:
dF: Freiheitsgrade; CFI: Comparative-Fit-Index;
RMSEA: Root-Mean-Square-Error of Approximation;
TLI: Tucker-Lewis-Index; SRMR: Standardized Root Mean Square Residual
Als Riskolernender gilt, wer mindestens eine Zugehörigkeit zu einer Risikogruppe aufweist.

Im *UDA-Assessment* lassen sich insgesamt akzeptable bis sehr gute Werte für die internen Konsistenzen beobachten (Tab. 5.34). Die Gruppe der Risikolernenden zeigt die besseren Werte für McDonalds-ω mit Ausnahme der *Herkunftsskala*. Hier unterscheiden sich die Werte jedoch nur geringfügig, so dass nicht von einer echten Benachteiligung der *Risikolernenden* durch das Testformat ausgegangen werden kann. Auffällig ist noch der schlechte McDonalds-ω-Wert für die *Sicherheitsskala* für die Gruppe der Nicht-Risikolernenden. Die Neuformulierung der negativ gepoolten Items hat demnach Auswirkungen auf die interne Konsistenz vor allem bezüglich der längeren Items der *Sicherheitsskala* (Durchschnitt: 10.3 Wörter pro Item) aber weniger auf die kürzeren Items der *Herkunftsskala* (Durchschnitt: 6.4 Wörter pro Item). Dies scheint in doppelter Hinsicht zu gelten: Während *Risikolernende* von den Adaptionen profitieren, scheinen diese die übrigen Schülerinnen und Schüler gemessen an den internen Konsistenzen zu benachteiligen. Die posiitv formulierten Items (*Rechtfertigungsskala*: durchschnittlich 10.9 Wörter pro Item; *Entwicklungsskala*: durchschnittlich 9.1 Wörter pro Item) zeigen weniger große Unterschiede hinsichtlich der internen Konsistenzen.

Im *Originalassessment* sind die internen Konsistenzen der Risikolernenden bezüglich der *Entwicklungsskala* (Durchschnitt: 10.6 Wörter pro Item) im Vergleich zur Gruppe der anderen ebenfalls besser (Tab. 5.34). Dies gilt auch für die *Rechtfertigungsskala* (Durchschnitt: 12.9 Wörter pro Item). Allerdings sind hier die Koeffizienten nur mäßig bzw. am Rande der Interpretierbarkeit. Die *Sicherheitsskala* (Durchschnitt: 12 Wörter pro Item) weist für beide Gruppen sehr schlechte interne Konsistenzen auf. Die *Herkunftsskala* (Durchschnitt: 6.4 Wörter pro Item) weist im Vergleich der Gruppen aber auch im Vergleich der Assessmentformen die besseren Werte auf.

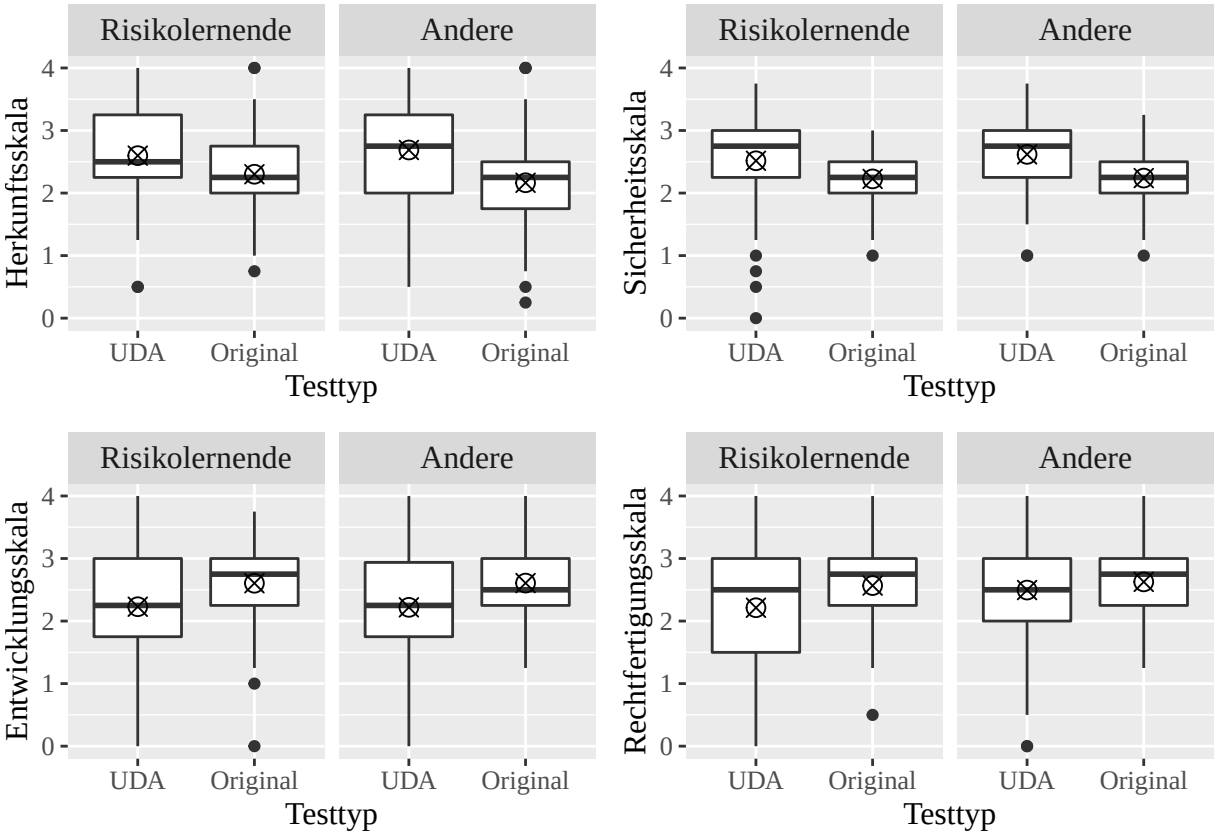


Abbildung 5.9: Boxplots zu den Mittelwerten der NOS-Kurzskalen beider Assessmentversionen aufgetrennt nach Risikolernenden und Anderen. Die Mittelwerte sind markiert.

Tabelle 5.36: T-Tests zu den Mittelwerten der Skalen und über beide Assessmentversion getrennt nach Risikolernenden und Anderen.

	Risikolernende			Andere		
NOS-Skala	UDA	Orig.	p	UDA	Orig.	p
Herkunftsskala	2.6	2.3	.05	2.69	2.16	.000
Sicherheitsskala	2.52	2.23	.01	2.62	2.24	.000
Entwicklungsskala	2.23	2.6	.01	2.22	2.61	.000
Rechtfertigungsskala	2.21	2.57	.05	2.5	2.63	n.s

Anmerkung:
UDA: UDA-Assessment; Orig.: Originalassessment
Umkodierung der negativen Skalen ist erfolgt.

In einem zweiten Schritt wurden die Mittelwerte jeder Skala verglichen. Durch die unterstützte (partial-)skalare Messinvarianz wird der Vergleich der Kurzskalenmittelwerte möglich. Die Boxplots für jede NOS-Skala getrennt nach Assessmentform und Risikolernenden und Anderen zeigen einheitliche Muster (Abb. 5.9). Der Interquartilbereich ist in allen Vergleichen für das UDA-Assessment größer. Weder besteht im UDA-Assessment (Herkunftsskala: $t(129) = 0.71$, p = n.s.; Sicherheitsskala: $t(98) = 0.98$, p = n.s.; Entwicklungsskala: $t(112)$ = -0.07, p = n.s.; Rechtfertigungsskala: $t(100) = 1.9$, p = n.s.), noch im Originalassessment (Herkunftsskala: $t(162)$ = -1.18, p = n.s.; Sicherheitsskala: $t(161)$

= 0.17, p = n.s.; Entwicklungsskala: $t(112) = 0.02$, p = n.s.; Rechtfertigungsskala: $t(100)$ = 0.67, p = n.s.) ein statistisch signifikanter Unterschied zwischen den Mittelwerten der Gruppen.

Auf Basis der internen Konsistenzen der Kurzskalen beider Assessments (die Analyse erfolgt mit dem R-Paket `psych` (Revelle, 2018) in Kombination mit `semTools` (Jorgensen et al., 2018)) lässt sich entsprechend der Einschränkungen ein *Differential Boost* nach Elliott et al. (2018) beobachten (Tab. 5.34). Allerdings stoßen die geringeren Werte für die übrigen Schülerinnen und Schüler auf. Nach Phillips (1994) hingegen kann auf der Basis der Mittelwerte kein *Differential Boost* festgestellt werden.

5.2.5 Entwicklung der interindividuellen NOS-Konzepte

Für die interindividuelle änderung der NOS-Konzepte wurden die *Lesefähigkeit*, die *Intelligenz* und der *sozioökonomische Status* sowie ein möglicher *sonderpädagogischer Unterstützungsbedarf* ausgewählt. Diese Merkmale sagen klassischer Weise interindividuelle änderungen hervor (Nehring et al., 2015). Außerdem wird der Effekt der Lernumgebung bestimmt. Aufgrund der geringen Anzahl von Schülerinnen und Schülern mit *sonderpädagogischem Unterstützungsbedarf* in der Stichprobe wird hierfür auch ein Signifikanzniveau von $p < .1$ akzeptiert.

Herkunftsskala

Für die *Herkunftsskala* lassen sich für das größere UDA-Assessments (Messzeitpunkt 1: M = 2.08; Messzeitpunkt 2: M = 2.37) und das Originalassessment (Messzeitpunkt 1: M = 2.64; Messzeitpunkt 2: M = 2.68) signifikante, latente Mittelwertsänderungen beobachten (Abb. 5.10), wobei der Unterschied im UDA-Assessment größer ist als im Originalassessment. Der einflussreichste Prädiktor für die *Herkunftsskala* zum zweiten Messzeitpunkt stellt der Wert *Herkunftsskala* zum Messzeitpunkt 1 dar. Sonst stellt nur noch die *Lesefähigkeit* im Originalassessment einen kleinen signifikanten Einfluss dar.

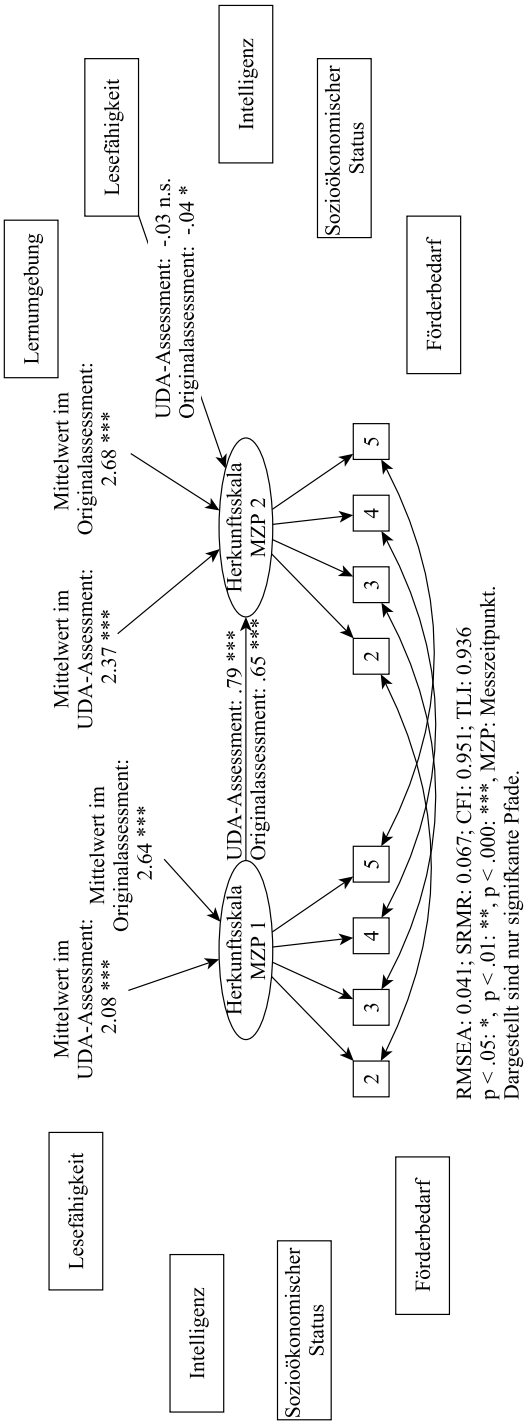


Abbildung 5.10: Vereinfachtes Multi-Group-Panelmodell zu den interindividuellen änderungen der Herkunftsskala unter Berücksichtigung der Lernendenmerkmale. Es sind nur signifkante Pfade dargestellt.

Sicherheitsskala

In ähnlicher Weise lässt sich auch die *Sicherheitsskala* beschreiben. Auch hier finden signifikante Mittelwertsänderungen im Originalassessment statt (Messzeitpunkt 1: M = 3.23; Messzeitpunkt 2: M = 3.34), die aber im UDA-Assessment (Messzeitpunkt 1: M = 2.37; Messzeitpunkt 2: M = 2.94) größer ausfallen (Abb. 5.11). Ebenso erweist sich

das Konstrukt als relativ stabil über die Intervention. Auch für die Skala regressiert die *Lesefähigkeit* signifkant auf das Konstrukt zum zweiten Messzeitpunkt. Außerdem beeinflusst die Lesefähigkeit das Abschneiden im Originalassessment zum ersten Messzeitpunkt. Weiterhin werden zum ersten Messzeitpunkt im Originalassessment *Schülerinnen und Schüler ohne sonderpädagogischen Unterstützungsbedarf* nicht signifikant bevorteilt. Im UDA-Assessment erreichen *Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf* hingegen höhere Werte. Zum zweiten Messzeitpunkt schneiden *Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf* im Originalassessment weiter schlechter ab, während der *sonderpädagogische Unterstützungsbedarf* im UDA-Assessment keine statistisch signifikante Rolle mehr spielt.

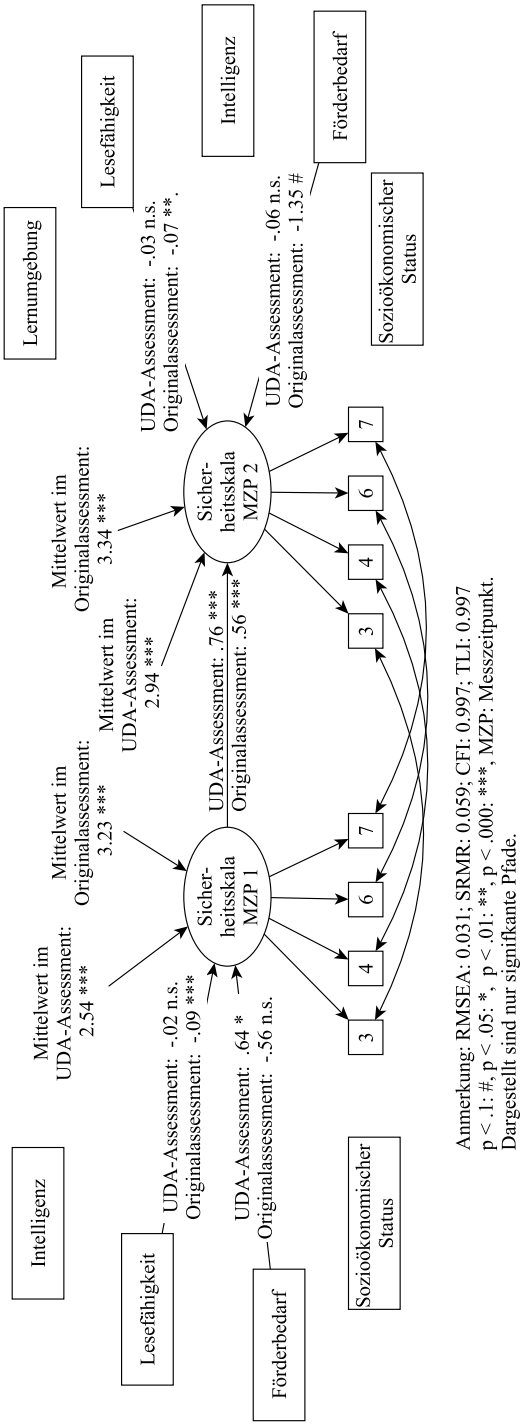


Abbildung 5.11: Vereinfachtes Multi-Group-Panelmodell zu den interindividuellen änderungen der Sicherheitsskala unter Berücksichtigung der Lernendenmerkmale. Es sind nur signifkante Pfade dargestellt.

Entwicklungsskala

Auch das NOS-Konstrukt *Entwicklung* erweist sich über beide Messzeitpunkte als relativ stabil. Die latenten Mittelwerte ändern sich auch für diese Skala in signifikanter Weise (Messzeitpunkt 1: UDA-Assessment M = 3.27, Originalassessment M = 3.60; Messzeitpunkt 2: UDA-Assessment M = 3.59, Originalassessment M = 3.69) (Abb. 5.12). Zum

ersten Messzeitpunkt hat die *Lesefähigkeit* einen leichten Einfluss für das Abschneiden im UDA-Assessment. Der *sonderpädagogische Unterstützungsbedarf* wirkt sich jedoch positiv auf das Abschneiden im Originalassessment aus. Zum zweiten Messzeitpunkt besteht zudem eine signifikante Regression der *Lesefähigkeit* im Originalassessment (Abb. 5.12).

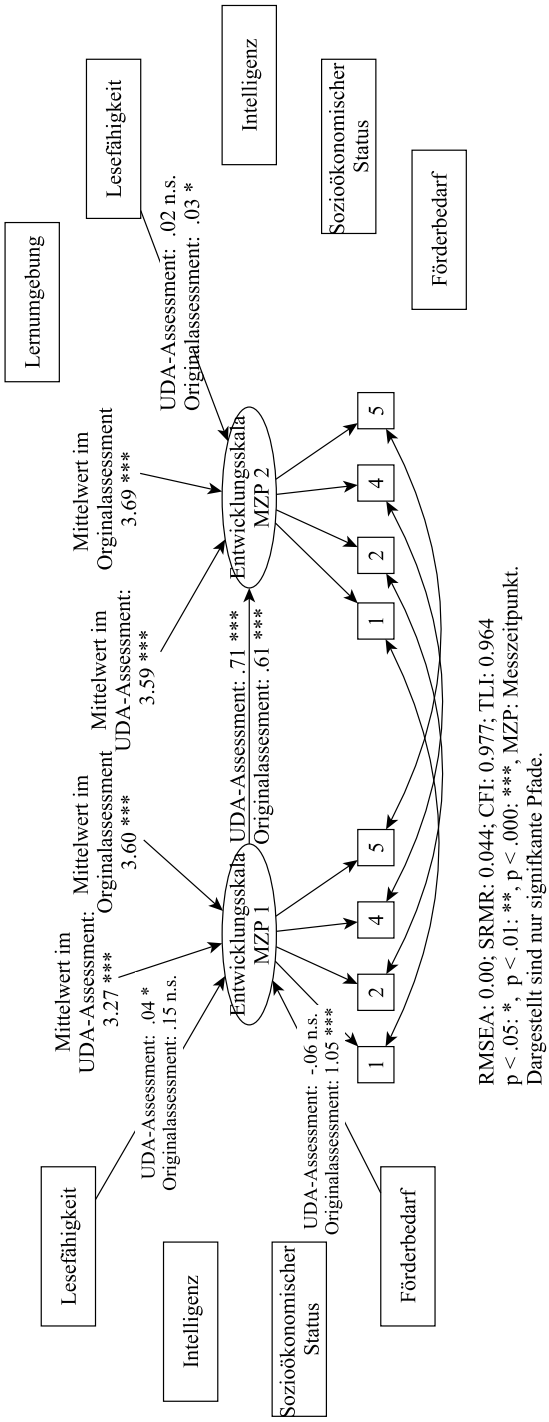


Abbildung 5.12: Vereinfachtes Multi-Group-Panelmodell zu den interindividuellen änderungen der Entwicklungsskala unter Berücksichtigung der Lernendenmerkmale. Es sind nur signifkante Pfade dargestellt.

Rechtfertigungsskala

Auch in der von den Lernumgebungen im Besonderen adressierten NOS-Dimension *Rechtfertigung* steigen die Mittelwerte signifikant (Messzeitpunkt 1: UDA-Assessment M = 3.44, Originalassessment M = 3.68; Messzeitpunkt 2: UDA-Assessment M = 3.67, Originalassessment M = 3.75) (Abb. 5.13). Zum ersten Messzeitpunkt sind die Regressionen zur *Intelligenz* und der *Lesefähigkeit* zum latenten Konstrukt signifikant. Die Regression der Konstrukte untereinander sind von allen Vergleichen die größten. Zum zweiten Messzeitpunkt zeigen *Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf* im Originalassessment geringere Werte als solche ohne.

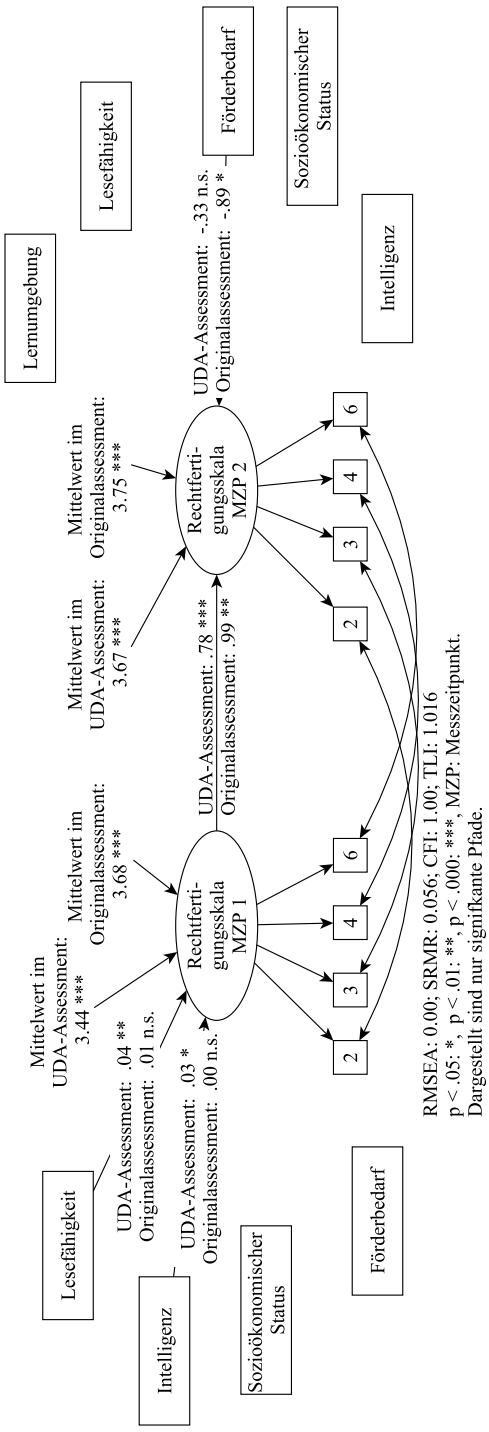


Abbildung 5.13: Vereinfachtes Multi-Group-Panelmodell zu den interindividuellen änderungen der Rechtfertigungsskala unter Berücksichtigung der Lernendenmerkmale. Es sind nur signifkante Pfade dargestellt.

Die Tabellen 5.37 und 5.38 zeigen eine Zusammenfassung der Modelle und der Effekttstärken (Cohens d). Hierbei ist auffällig, dass die Effektstärken für das UDA-Assessment als klein zu beschreiben sind. Diese sind jedoch immer größer als im Orginalassessment. Hier ist nur der Effekt für die Entwicklungsskala relevant.

5.2.6 Entwicklung der intraindividuellen NOS-Konzepte

Bezüglich eines *Lehrens und Lernens für alle* werden im Folgenden die intraindividuellen änderungen beschrieben. Da außerdem bisweilen kaum ein Effekt der *Lesefähigkeit*, des *Geschlechts*, des *sonderpädagogischen Unterstützungsbedarfs*, des *sozioökonomischen Statuses* sowie der *Intelligenz* auf die Werte der NOS-Skalen beobachtet werden konnte, wird im Weiteren auf die Analyse zu diesen Variablen verzichtet. Stattdessen werden mögliche Effekte der *wahrgenommenen kognitiven Aktivierung* (Tab. 7.15; (Fauth et al., 2014)) und des *wahrgenommenen Nutzens des E-Books* (Tab. 7.16; (Williamson Sprague & Dahl, 2010)) sowie der *Lernumgebung* auf die intraindividuellen änderungen bestimmt.

über Varianzanalysen zeigt sich, dass auf manifester Ebene nur die Art der Testform nicht aber die Art der Lernumgebung einen signifikanten Einfluss auf die Daten zum *wahrgenommenen Nutzen des E-Books* aufweist ($F(1{,}299) = 170.69$, $p < .000$). Eine Post-hoc-Analyse über den Tukeytest zeigt Mittelwertsunterschiede zwischen dem UDA- (M = 4.31) und dem Originalassessment (M = 3.10) auf einem Signifikanzniveau von $p < .000$. Entsprechend ist der *wahrgenommene Nutzen der E-Books* im UDA-Assesment höher. Im Originalassessment hingegen ist die *wahrgenommene kognitive Aktivierung* höher ($F(1{,}281) = 39.26$, $p < .000$; Tukeytest: Originalassessment: M = 3.96, UDA-Assessment: M = 3.32, $p < .000$).

Für die *Herkunftsskala* zeigt sich, dass insbesondere zum ersten Messzeitpunkt für die UDA-Version die Antworten der Schülerinnen und Schüler geringer sind als zum zweiten Messzeitpunkt (Abb. 5.14). Die Mittelwertsgerade im UDA-Assessment ($y = 0.306x + 2.078$; ; $M_1 = 2.38$, $M_2 = 2.69$) zeigt eine größere Steigung im Kontrast zum Originalassessment ($y = 0.031x + 2.674$; ; $M_1 = 2.71$, $M_2 = 2.74$) (Kovarianz im UDA-Assessment: $\phi = .11$, $p < .01$; Kovarianz im Originalassessment: $\phi = .06$, $p < .05$). Für beide Assessmentversionen lässt sich eine größere Varianz in den Daten zum zweiten Messzeitpunkt beobachten. Kategorisiert man die Differenzen nach größer und kleiner Null erzielen 131 (74.86 %) Lernende im UDA-Assessment und 100 (60.61 %) Lernende im Originalassessment einen höheren Wert zum zweiten Messzeitpunkt als zum ersten (Tab. 5.39). In beiden Assessments zeigen aber auch eine Reihe von Schülerinnen und Schülern schlechtere Werte zum zweiten Messzeitpunkt (UDA-Assessment: n = 40 (22.86 %); Originalassessment: n = 62 (37.58 %)).

Mit Bezug zur *Sicherheitsskala* lässt sich im Originalassessment ein ähnliches Bild verzeichnen (Abb. 5.14). Auch hier steigt die Varianz zum zweiten Messzeitpunkt an. Die Mittelwertsgerade zeigt jeweils eine leichte Steigung (UDA-Assessment: $y = 0.401 * x + 2.475$; $M_1 = 2.88$, $M_2 = 3.28$; $\phi = .17$, p < .05; Originalassessment: $y = 0.150 * x + 3.272$; $M_1 = 3.42$, $M_2 = 3.57$; $\phi = .10$, $p < .05$). Im Vergleich sind die Mittelwerte des UDA-Assessments erneut niedriger. Auch vergrößert sich die Varianz der Daten zum zweiten Messzeitpunkt. Mit Blick auf die Verteilung der Differenzen weisen 173 (98.29 %) Lernende im UDA-Assessment und 126 (76.36 %) im Originalassessment eine Differenz, die größer als Null ist auf (Tab. 5.39). Im Originalassessment zeigen 34 (20.61 %) der Lernenden einen schlechteren Wert zum zweiten Messzeitpunkt. Dies trifft auf keinen Lernenden im UDA-Assessment zu. Die Kovarianz der Second-order-Konstrukte ist in beiden Assessments positiv und signifkant (Kovarianz im UDA-Assessment: $\phi = .17$, $p < .05$; Kovarianz im Originalassessment: $\phi = .1$, $p < .05$).

Die longitudinal Plots der jeweiligen Assessments der *Entwicklungsskala* fügen sich in

die bisherigen Beschreibungen (Abb. 5.14). Im UDA-Assessment steigt der Durchschnitt vom ersten zum zweiten Messzeitpunkt in ähnlicher Weise wie zuvor an ($y = 0.351 * x + 3.204$; $M_1 = 3.56$, $M_2 = 3.91$). Auch im Originalssessment steigt der Durchschnitt leicht an ($y = 0.086 * x + 3.556$; $M_1 = 3.69$, $M_2 = 3.76$). Während die Varianzen im UDA-Assesment zu beiden Messzeitpunkten vergleichbar großsind, steigt die Varianz zum zweiten Messzeitpunkt im Orginalassessment. Die Kovarianz zwischen dem longitudinalen Mittelwert und der änderungsrate ist im UDA-Assesment positiv aber nicht signifikant ($\phi = .05$, $p =$ n.s.), im Originalassessment hingegen schon ($\phi = .072$, $p < .01$) Die Verteilung der Differenzen fällt vergleichbar aus: 174 (99.43 %) Lernende im UDA-Assessment und 114 (69.09 %) im Originalassessment weisen eine Differenz, die größer als Null ist auf (Tab. 5.39). Auch zu dieser Skala zeigen im Originalassessment 48 (29.09 %) der Lernenden einen schlechteren Wert zum zweiten Messzeitpunkt. Dies trifft auf niemanden im UDA-Assessment zu.

Die longitudinal Plots der jeweiligen Assessments der *Rechtfertigungsskala* schließen sich dem bisherigen Bild an (Abb. 5.14). Im UDA-Assessment steigt der Durchschnitt vom ersten zum zweiten Messzeitpunkt ($y = 0.229 * x + 3.376$; $M_1 = 3.61$, $M_2 = 3.83$). Gleiches gilt für das Originalassessment ($y = 0.076 * x + 3.61$; $M_1 = 3.69$, $M_2 = 3.76$). Wie bei der *Entwicklungsskala* sind die Varianzen im UDA-Assesment zu beiden Messzeitpunkten vergleichbar groß. Ebenso wie zuvor steigt die Varianz zum zweiten Messzeitpunkt im Orginalassessment. Die Verteilung der Differenzen fällt vergleichbar aus: 174 (99.43 %) Lernende im UDA-Assessment und 115 (69.70 %) im Originalassessment weisen eine Diffenrenz auf, die größer als Null ist (Tab. 5.39). Die Kovarianz ist statistisch nicht signifikant für die Assessments (UDA-Assessment: $\phi = .02$, $p <$ n.s.; Originalassessment: $\phi = .05$, $p <$ n.s.). Hier zeigen 47 (28.48 %) der Lernenden einen schlechteren Wert zum zweiten Messzeitpunkt. Abermals zeigt kein Lernender im UDA-Assessment einen schlechteren Wert.

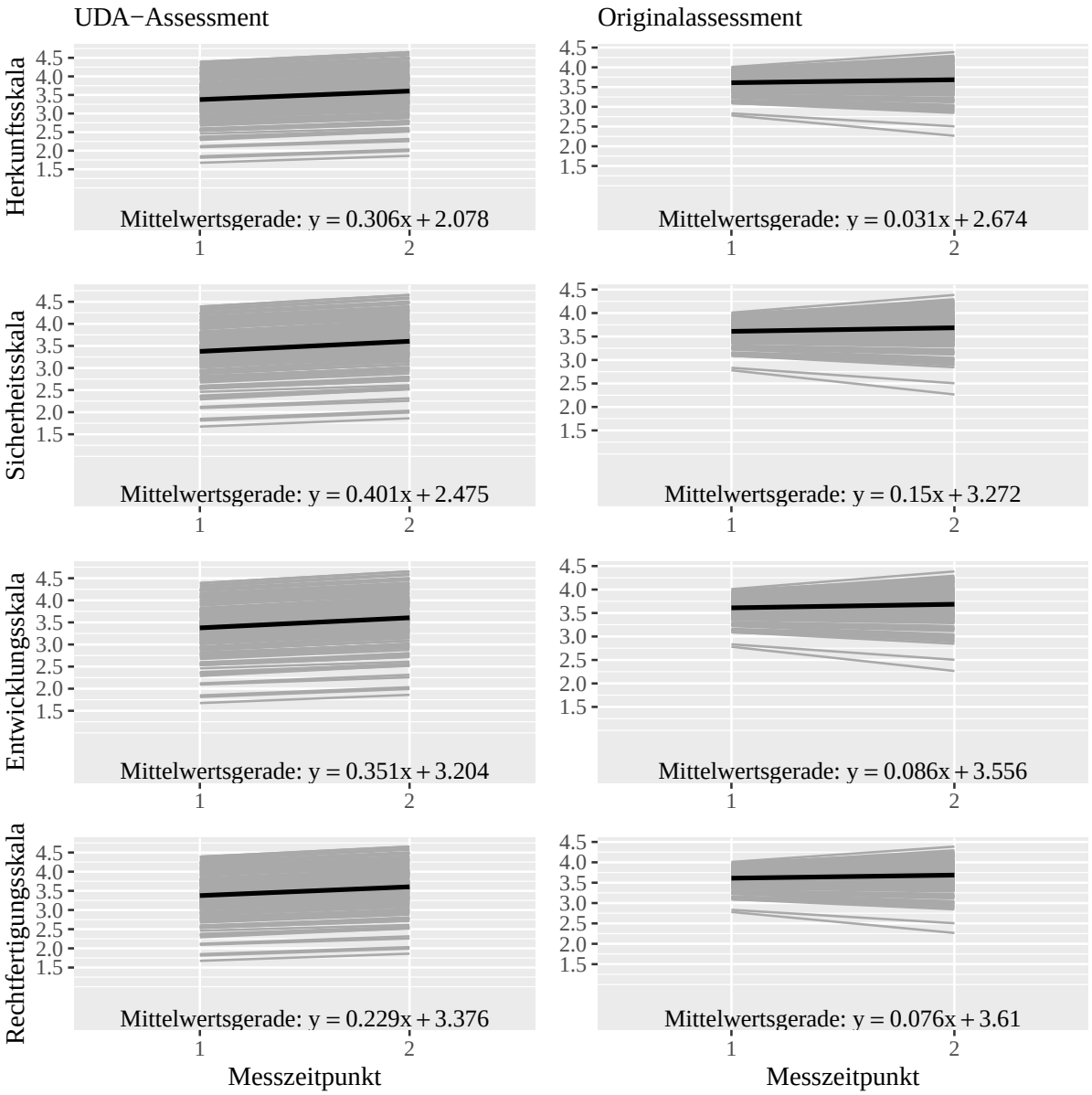


Abbildung 5.14: Longitudinal Plots zu Trajektorien aller Schülerinnen und Schüler zu den Skalen aus beiden Assessments (UDA-Assessment n = 175; Originalassessment= 165). Der Durchschnitt ist schwarz dargestellt.

In einem zweiten Schritt wird der Einfluss der wahrgenommenen *kognitiven Aktivierung*, des *persönlichen Nutzen* des E-Books und die Art der *Lernumgebung* auf die Trajektoren bestimmt. Für die *Herkunftsskala* (CFI = 0.976, RMSEA = 0.011) ergibt sich, dass die Lerngruppe signifikant auf die longitudinalen Mittelwert regressiert (Abb. 5.15). Dies fügt sich in das Bild der longitudinalen Plots (Abb. 5.14) und stellt eine Heterogenität in der Stichprobe mit Bezug zur *UDL-Lernumgebung* dar (UDA-Assessment: β = -.36, $p < .01$; Originalassessment: β = .05, p = n.s.). Dies lässt sich für das Originalassessment nicht beobachten; hier besteht kein signifikanter Einfluss der Lernumgebung auf die Mittelwerte. Bezüglich der *änderungsrate* führt eine höhere *wahrgenommene kognitive Aktivierung* zu einer niedrigeren änderungsrate im Originalassessment (UDA-Assessment: β = -.01, p = n.s.; Originalassessment: β = -.21, $p < .01$). Die Kovarianz zwischen den Second-Order-Konstrukten ist in beiden Fällen positiv signifikant (UDA-Assessment: ϕ = .112, $p < .01$;

Originalassessment: $\phi = .06$, $p < .05$).

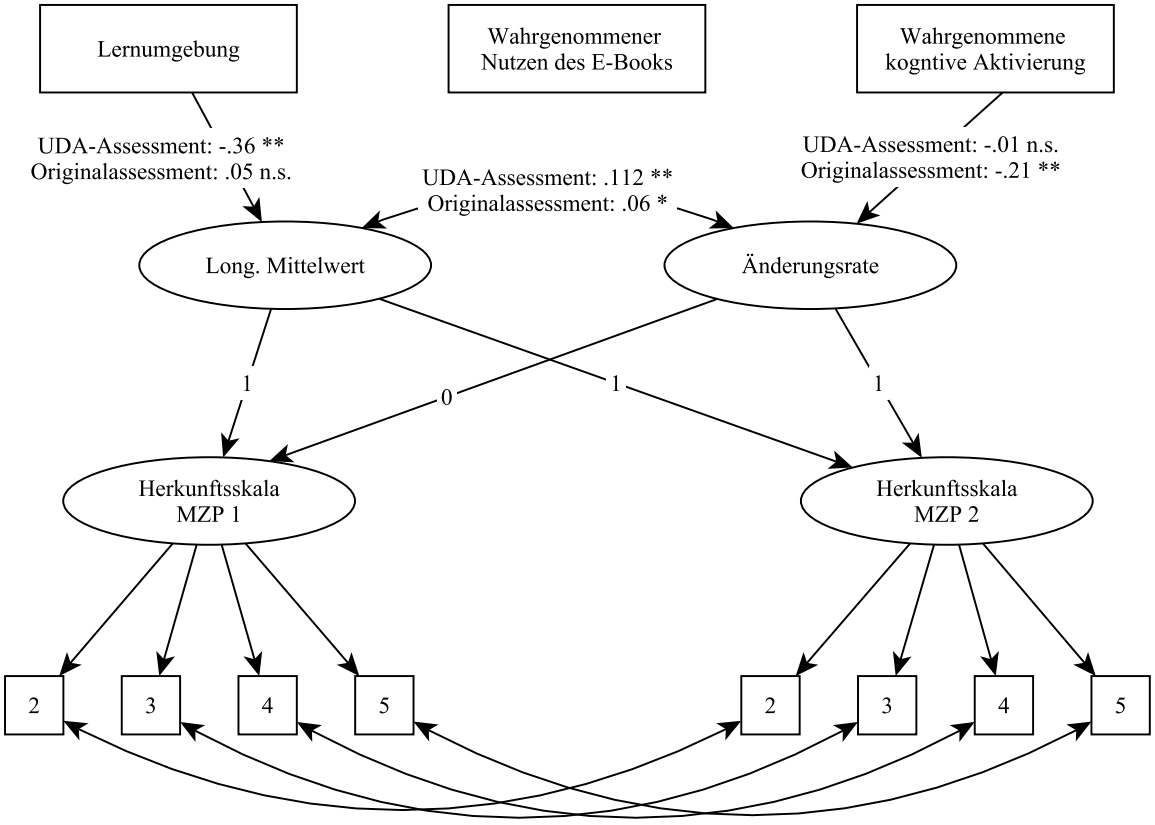


Abbildung 5.15: Vereinfachtes latentes Wachstumsmodell zur Herkunftsskala aus beiden Assessments unter Berücksichtigung der wahrgenommenen kognitiven Aktivierung und des persönlichen Nutzens des E-Books sowie der Lernumgebung.

Im Modell für die *Sicherheitsskala* (CFI = 0.973, RMSEA = 0.046) ist lediglich die Kovarianz in beiden Assessments zwischen dem longitudinalen Mittelwert und der änderungsrate signifikant (Abb. 5.16). Weder die *Lernumgebung* noch der *wahrgenommene Nutzen des E-Books* oder die *kognitive Aktivierung* sind statistisch relevant für die Entwicklungen. Die Kovarianz ist in beiden Fällen positiv signifkant (UDA-Assessment: $\phi = .165$, $p < .05$; Originalassessment: $\phi = .103$, $p < .05$).

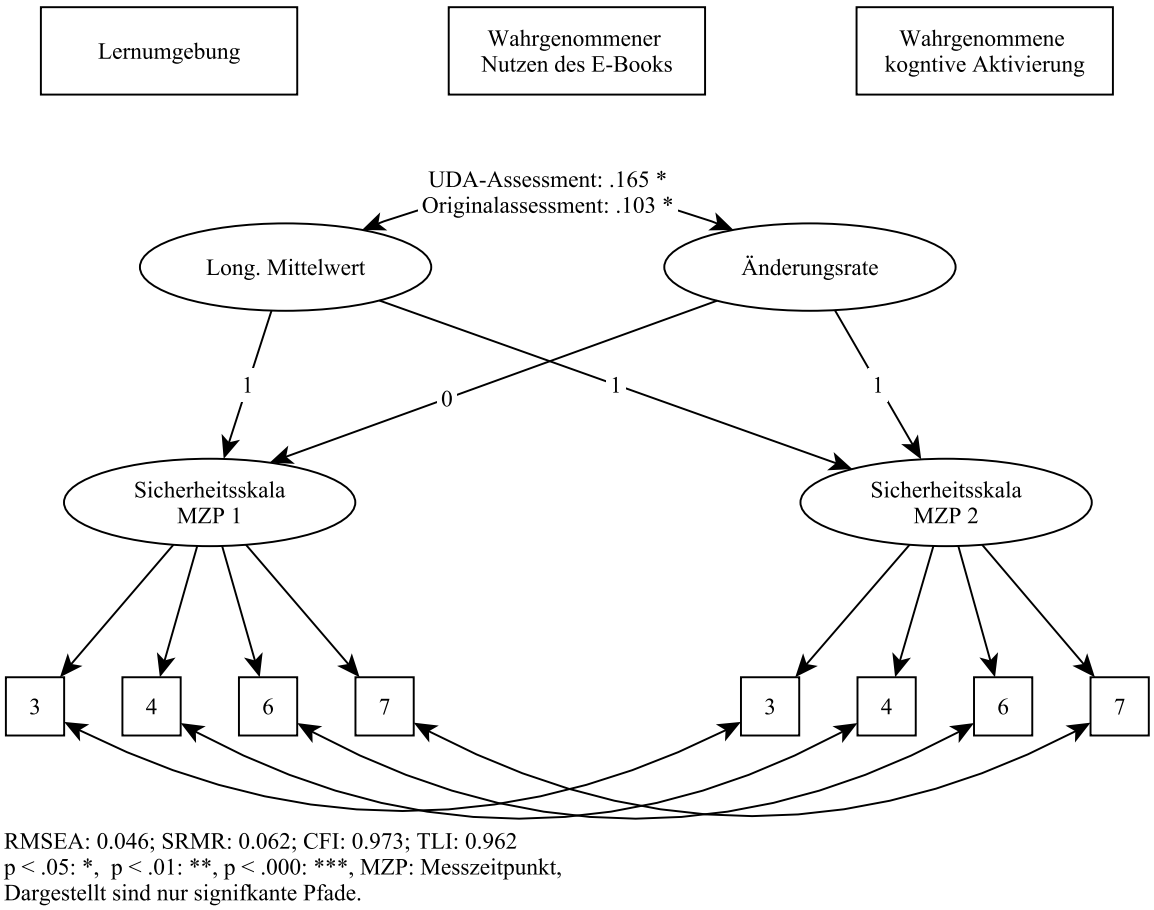


Abbildung 5.16: Vereinfachtes latentes Wachstumsmodell zur Sicherheitsskala aus beiden Assessments unter Berücksichtigung der wahrgenommenen kognitiven Aktivierung und des persönlichen Nutzens des E-Books sowie der Lernumgebung.

Der longitudinale Mittelwert (UDA-Assessment: β = -.10, p = n.s.; Originalassessment: β = .18, $p < .05$) und die änderungsrate (UDA-Assessment: β = .01, p = n.s.; Originalassessment: β = .18, $p < .05$) der *Entwicklungsskala* (CFI = 0.998, RMSEA = 0.017) werden nur von der *wahrgenommenen kognitiven Aktivierung* signifikant im Originalassessment regressiert (Abb. 5.17). Sowohl für den longitudinalen Mittelwert als auch für dessen änderungsrate führt eine höhere *wahrgenommene kognitive Aktivierung* auch zu höheren Werten auf der Skala. Die Kovarianz zwischen beiden latenten Konstrukten zweiter Ordnung ist gering und nur im Originalassessment signifikant (UDA-Assessment: ϕ = .05, p = n.s.; Originalassessment: ϕ = .07, $p < .05$).

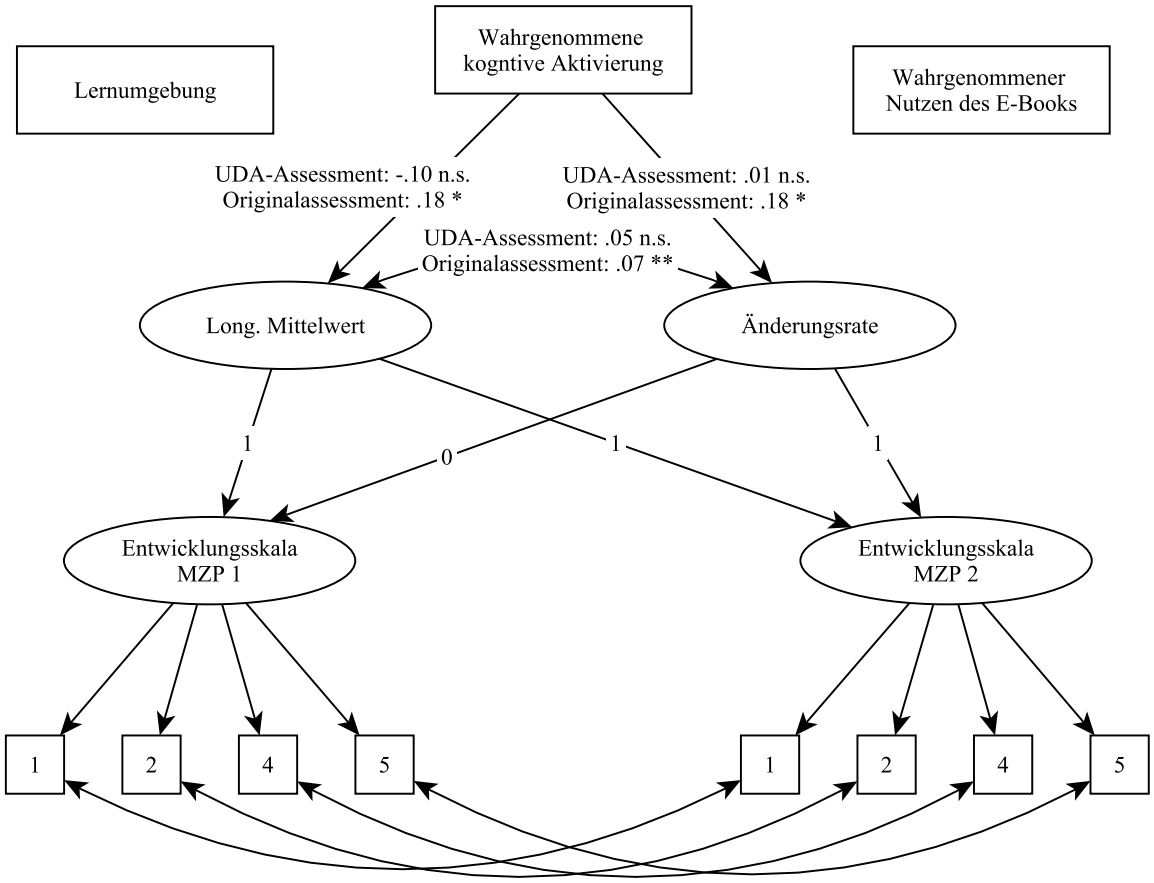


Abbildung 5.17: Vereinfachtes latentes Wachstumsmodell zur Entwicklungsskala aus beiden Assessments unter Berücksichtigung der wahrgenommenen kognitiven Aktivierung und des persönlichen Nutzens des E-Books sowie der Lernumgebung.

ähnliches gilt für den longitudinalen Mittelwert und die änderungsrate der *Rechtfertigungsskala* (CFI = 1.00, RMSEA = 0.00). Diese werden durch die *wahrgenommene kognitiven Aktivierung* signifikant im Originalassessment regressiert (long. Mittelwert: UDA-Assessment: β = -.04, p = n.s.; Originalassessment: β = .16, $p < .05$; änderungsrate: UDA-Assessment: β = -.03, p = n.s.; Originalassessment: β = .17, $p < .05$) (Abb. 5.18). Demnach führt auch hier eine *höhere kognitive Aktivierung* zu einer Erhöhung des longitudinalen Mittelwerts und dessen änderungsrate. Zusätzlich weist das Modell für die UDL-Lernumgebung (UDA-Assessment: β = -.28, $p < .05$; Originalassessment: β = -.04, p = n.s.; eine geringere und signifikante änderungsrate auf. Unter Hinzunahme der Kontrollvariablen ist die Kovarianz zwischen dem logitudinalen Mittelwert und der änderungsrate im Modell für die *Rechtfertigungsskala* gering und statistisch nicht mehr signifkant (UDA-Assessment: ϕ = .05, p = n.s.; Originalassessment: ϕ = .05, p = n.s.).

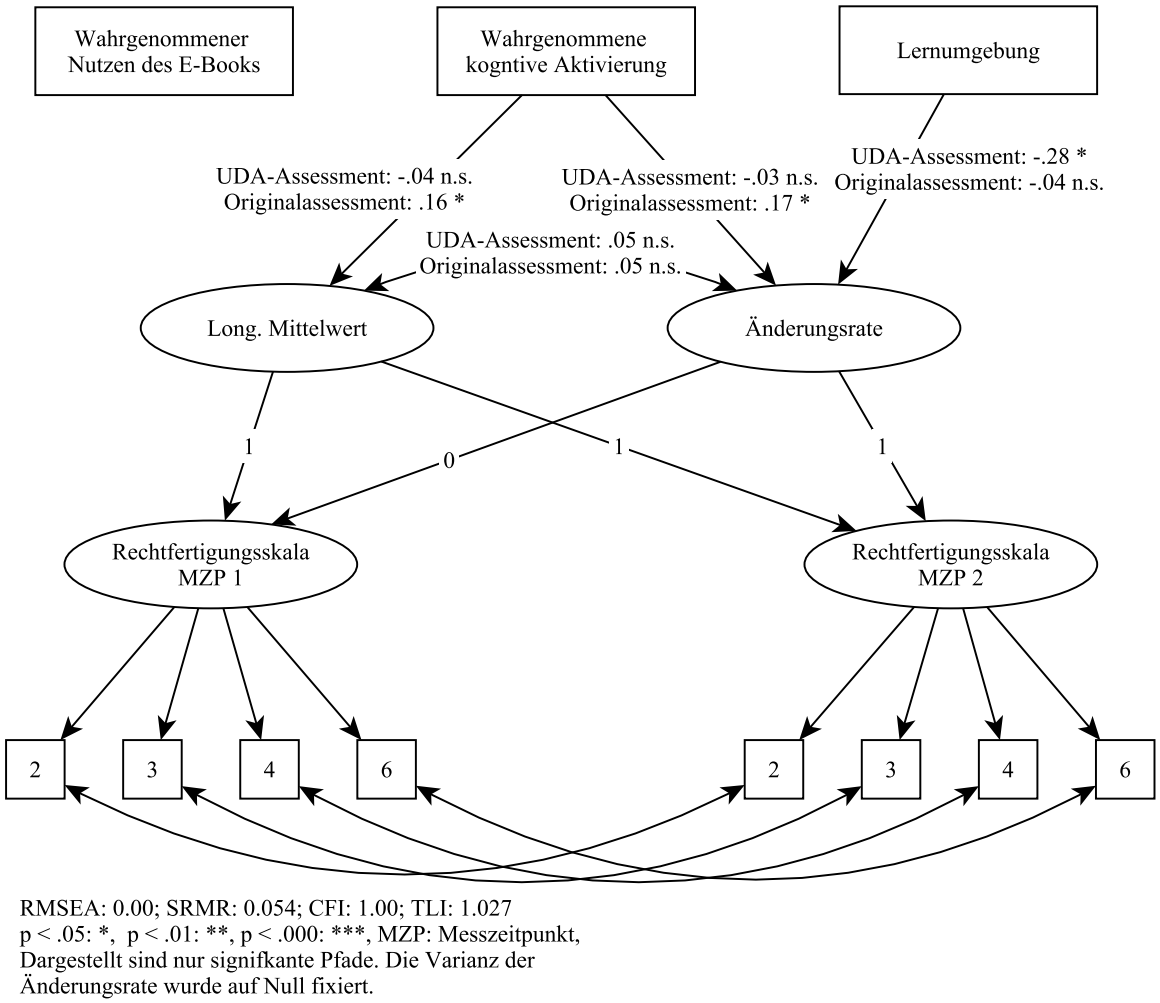


Abbildung 5.18: Vereinfachtes latentes Wachstumsmodell zur Rechtfertigungsskala aus beiden Assessments unter Berücksichtigung der wahrgenommenen kognitiven Aktivierung und des persönlichen Nutzens des E-Books sowie der Lernumgebung.

Die Tabellen 5.40 und 5.41 zeigen eine Zusammfassung zum Einfluss der Kovariaten in den latenten Wachstumsmodellen.

Tabelle 5.37: Einfluss der Kovariaten im UDA-Assessment in den Panelmodellen zum jeweiligen Messzeitpunkt (MZP), Mittelwerte und Effektstärken in Cohens d.

	MZP 1		MZP 2		
	Koeffizient	p-Wert	Koeffizient	p-Wert	
Herkunft					
MZP 1	—	—	0.79	<0.05	—
Lernumgebung	—	—	0.26	0.13	—
Lesefäigkeit	0	0.86	-0.02	0.15	—
Intelligenz	0	0.95	0	0.87	—
SES	-0.03	0.46	0.06	0.32	—
SU	0.44	0.11	0.29	0.37	—
Mittelwert	—	2.08	—	2.37	—
Effektstärke	—	—	—	—	0.37
Sicherheit					
MZP 1	—	—	0.75	<0.05	—
Lernumgebung	—	—	0.26	0.21	—
Lesefäigkeit	-0.02	0.33	-0.03	0.15	—
Intelligenz	-0.01	0.48	0.01	0.58	—
SES	0.01	0.84	0.09	0.21	—
SU	0.63	0.12	0.06	0.90	—
Mittelwert	—	2.37	—	2.94	—
Effektstärke	—	—	—	—	0.37
Entwicklung					
MZP 1	—	—	0.07	0.81	—
Lernumgebung	—	—	0.71	<0.05	—
Lesefäigkeit	0.07	0.65	0.09	0.53	—
Intelligenz	0.04	<0.05	0.01	0.76	—
SES	0.01	0.48	0	0.73	—
SU	-0.08	0.09	0.03	0.53	—
Mittelwert	—	3.27	—	3.59	—
Effektstärke	—	—	—	—	0.36
Rechtfertigung					
MZP 1	—	—	0.78	<0.05	—
Lernumgebung	—	—	-0.17	0.16	—
Lesefäigkeit	0.04	<0.05	0	0.72	—
Intelligenz	0.03	<0.05	-0.01	0.29	—
SES	0.01	0.76	-0.01	0.76	—
SU	0.47	0.18	-0.33	0.46	—
Mittelwert	—	3.44	—	3.67	—
Effektstärke	—	—	—	—	0.3

Anmerkung:
SU: sonderpädagogischen Unterstützungsbedarf.
SES: sozioökonomischer status

Tabelle 5.38: Einfluss der Kovariaten im Originalassessment in den Panelmodellen zum jeweiligen Messzeitpunkt (MZP), Mittelwerte und Effektstärken in Cohens d.

	MZP 1		MZP 2		
	Koeffizient	p-Wert	Koeffizient	p-Wert	
Herkunft					
MZP 1	—	—	0.65	<0.05	—
Lernumgebung	—	—	-0.05	0.70	—
Lesefäigkeit	-0.02	0.27	-0.04	<0.05	—
Intelligenz	0	0.83	0.01	0.14	—
SES	-0.02	0.69	0.01	0.70	—
SU	-0.08	0.82	-0.42	0.21	—
Mittelwert	—	2.64	—	2.68	—
Effektstärke	—	—	—	—	0.05
Sicherheit					
MZP 1	—	—	-0.01	<0.05	—
Lernumgebung	—	—	-0.52	0.65	—
Lesefäigkeit	0.32	<0.05	0.56	<0.05	—
Intelligenz	0.22	0.47	0.07	0.88	—
SES	-0.09	0.81	-0.07	0.69	—
SU	0.01	0.31	0	0.06	—
Mittelwert	—	3.23	—	3.34	—
Effektstärke	—	—	—	—	0.13
Entwicklung					
MZP 1	—	—	0.3	<0.05	—
Lernumgebung	—	—	0.08	<0.05	—
Lesefäigkeit	0.09	0.89	-0.54	0.86	—
Intelligenz	0.08	0.56	0.16	<0.05	—
SES	-0.54	0.21	0.3	0.62	—
SU	0.16	0.10	-0.02	0.70	—
Mittelwert	—	3.6	—	3.69	—
Effektstärke	—	—	—	—	0.15
Rechtfertigung					
MZP 1	—	—	0	<0.05	—
Lernumgebung	—	—	0.04	0.62	—
Lesefäigkeit	-0.01	0.69	-0.15	0.81	—
Intelligenz	-0.12	0.70	0.99	0.17	—
SES	-0.07	0.24	-0.06	0.57	—
SU	0.01	0.45	0	0.05	—
Mittelwert	—	3.64	—	3.74	—
Effektstärke	—	—	—	—	0.24

Anmerkung:
SU: sonderpädagogischen Unterstützungsbedarf.
SES: sozioökonomischer status

Tabelle 5.39: Anzahl der kategorisierten Differenzen aus den latenten Wachtumsmodellen.

	Absolute Verteilung			Relative Verteilung		
	< 0	> 0	NA	< 0	> 0	NA
UDA-Assessment (n = 175)						
Herkunft	n = 39	n =132	n = 4	22.86 %	74.86 %	2.29 %
Sicherheit	n = 0	n =173	n = 2	0.00 %	98.29 %	1.14 %
Entwicklung	n = 0	n =174	n = 1	0.00 %	99.43 %	0.57 %
Rechtfertigung	n = 0	n =174	n = 1	0.00 %	99.43 %	0.57 %
Originalassessment (n = 165)						
Herkunft	n = 62	n =100	n = 3	37.58 %	60.61 %	1.82 %
Sicherheit	n = 34	n =126	n = 5	20.61 %	76.36 %	3.03 %
Entwicklung	n = 48	n =114	n = 3	29.09 %	69.09 %	1.82 %
Rechtfertigung	n = 47	n =115	n = 3	28.48 %	69.70 %	1.82 %

Anmerkung:
NA wird vergeben, wenn keine Schätzung aufgrund von fehlenden Werten möglich war.

Tabelle 5.40: Einfluss der Kovariaten in den latenten Wachtumsmodellen im UDA-Assessment.

	änderungsrate			long. Mittelwert		
	Koeffizient	stand. Koeffizient	p-Wert	Koeffizient	stand. Koeffizient	p-Wert
Herkunft						
Kognitive Aktiverung	0.00	-0.03	0.97	0.09	0.86	0.39
Ebooknutzung	0.10	0.86	0.39	0.03	0.45	0.65
Lernumgebung	0.30	1.83	0.07	-0.36	-3.00	<0.05
Sicherheit						
Kognitive Aktiverung	-0.02	-0.15	0.88	0.11	0.74	0.46
Ebooknutzung	-0.06	-0.45	0.66	0.15	1.26	0.21
Lernumgebung	0.16	0.92	0.36	-0.05	-0.29	0.77
Entwicklung						
Kognitive Aktiverung	-0.10	-0.68	0.50	0.01	0.07	0.94
Ebooknutzung	0.04	0.33	0.74	-0.06	-0.62	0.54
Lernumgebung	0.07	0.45	0.65	0.03	0.22	0.82
Rechtfertigung						
Kognitive Aktiverung	-0.02	-0.23	0.82	-0.04	-0.34	0.73
Ebooknutzung	0.05	0.56	0.58	0.03	0.34	0.73
Lernumgebung	-0.28	-2.25	<0.05	0.13	0.92	0.36

Tabelle 5.41: Einfluss der Kovariaten in den latenten Wachtumsmodellen im Orginalassessment.

	änderungsrate			long. Mittelwert		
	Koeffizient	stand. Koeffizient	p-Wert	Koeffizient	stand. Koeffizient	p-Wert
Herkunft						
Kognitive Aktiverung	-0.20	-2.85	<0.05	0.10	1.35	0.18
Ebooknutzung	0.10	0.91	0.36	0.18	1.88	0.06
Lernumgebung	-0.14	-1.06	0.29	0.05	0.38	0.70
Sicherheit						
Kognitive Aktiverung	-0.14	-1.19	0.23	0.03	0.21	0.84
Ebooknutzung	0.32	1.70	0.09	0.16	1.15	0.25
Lernumgebung	0.03	0.19	0.85	0.03	0.19	0.85
Entwicklung						
Kognitive Aktiverung	0.18	2.41	<0.05	0.18	2.45	<0.05
Ebooknutzung	0.15	1.05	0.29	-0.02	-0.25	0.80
Lernumgebung	-0.05	-0.38	0.71	0.05	0.48	0.63
Rechtfertigung						
Kognitive Aktiverung	0.17	2.03	<0.05	0.16	2.63	<0.05
Ebooknutzung	-0.02	-0.24	0.81	0.12	1.40	0.16
Lernumgebung	-0.04	-0.39	0.69	0.08	0.83	0.41

5.2.7 Zusammenfassung der Hauptstudienergebnisse

Forschungsfrage 4: Inwiefern unterscheiden sich das adaptierte und das nicht-adaptierte Assessment in der Erfassung von NOS-Konzepten hinsichtlich der internen Konsistenz und der Skalenstruktur?

Die konfirmatorischen Faktorenanalysen (KFA) zeigen, dass je Skala und Assessmentform ein Globalfaktor angenommen werden kann. Allerdings sind die Inter-Item-Korrelationen im Originalassessment geringer als im UDA-Assessment mit Ausnahme der *Herkunftsskala*. So sind die internen Konsistenzen der gesamten *Herkunftsskala* im UDA-Assessment und im Originalassessment für die gesamte *Rechtfertigungsskala* schlechter im Vergleich zu den übrigen Skalen und befinden sich am Rande der Akzeptanz. Zum zweiten Messzeitpunkt steigen die internen Konsistenzen für alle Skalen in beiden Assessmentversionen an. Insgesamt erweisen sich die Skalen aus dem UDA-Assessment bereits zum ersten Messzeitpunkt mit Bezug auf McDonalds-ω als konsistenter. Entsprechend sind die Unterschiede zum zweiten Messzeitpunkt weniger gravierend.

Forschungsfrage 5: Inwiefern unterscheiden sich das adaptierte und das nicht-adaptierte Assessment in der Erfassung von NOS-Konzepten hinsichtlich der longitudinalen Messinvarianz und der instruktionalen Sensitivität?

Für die Vollskalen lässt sich in drei von vier Fällen nur metrische MI feststellen. Dies begründet sich einerseits in der Anzahl an Items und der daraus resultierenden Varianz, die andererseits in einem ungünstigen Verhältnis zur Stichprobengröße und den zu schätzenden Parametern steht. Da die Instrumente in dieser Studie aber nicht für sich evaluiert, sondern für die Messung von änderungen von NOS-Konzepten verwendet werden, wurden sie auf instruktionale Sensitivität hin untersucht. über die Analyse der Faktorladungen und über manifeste T-Tests wurden Items identifiziert, die sich signifikant zwischen den Messzeitpunkten ändern und damit als instruktional sensitiv beschrieben werden können. Auf dieser Basis wurden Skalen einheitlicher Größe mit vier Items gebildet, um die änderungen durch die Intervention möglichst genau abzubilden und die formale Validität zu erhöhen. Außerdem steigen die internen Konsistenzen aller Kurzskalen mit Ausnahme der Rechtfertigungsskala im Originalassessment auf gute bis sehr gute Werte. Die MI-Prüfung mit den Kurzskalen zeigt zum Teil sehr gute Werte der angeführten Fitindezes (CFI, TLI, RMSEA und SRMR), wobei lediglich für die *Rechtfertigungsskala* die volle skalare MI und darüber hinaus auch noch die strikte MI angenommen werden kann. Allerdings unterstützen die anderen Skalen in jedem Fall mindestens die partial-skalare MI. Wenn die Beschränkungen aufgehoben werden mussten, dann deshalb, weil es zum ersten Messzeitpunkt gravierende Unterschiede in den Beantwortungen der Items gab. Dies ändert sich zum zweiten Messzeitpunkt hin. Auffällig ist noch, dass im MIMIC-Ansatz die Lernumgebung jeweils nur auf bestimmte Items jedoch nicht auf den latenten Faktor signifikant regressiert. Dies spricht für die Gleichwertigkeit beider Lernumgebungen.

Forschungsfrage 6: Inwiefern unterscheiden sich das adaptierte und das nicht-adaptierte Assessment in der Erfassung von NOS-Konzepten hinsichtlich der Gruppenabhängigkeit (Lernendenmerkmale) von Items (Testzugänglichkeit)?

Für die Untersuchung der Gruppenabhängigkeit und damit Testzugänglichkeit der Items wurden Cut-off-Werte definiert. Auf Basis dieser Werte für den *sozioökonomischen Status*, der *Lesefähigkeit* und der *Intelligenz* wurden Gruppen gebildet. Zusätzlich wurde noch die Gruppenabhängigkeit mit Bezug zum *Geschlecht* analysiert. Der *sonderpädagogische Unterstützungsbedarf* wurde ausgelassen aufgrund der zu kleinen Stichprobe. In der statistischen Beschreibung zeigt sich aber, dass sich ein diagnostizierter sowie ein möglicher sonderpädagogischer Unterstützungsbedarf durch die *Lesefähigkeit* und die *Intelligenz* abbilden lässt. Die grafische MI-Prüfungen der DIF-Analysen lassen schließen, dass einige Items aus dem Originalassessment Gruppenabhängigkeit zeigen. Außerdem fällt auf, dass die Konfidenzintervalle größer sind als im UDA-Assessment. Damit erweist sich das Originalassessment als weniger testzugänglich als das UDA-Assessment, weil die Probandencharakteristiken, und hier vor allem der *sozioökonomische Status*, das Antwortverhalten beeinflussen.

Forschungsfrage 7: Inwiefern unterscheiden sich das adaptierte und das nicht-adaptierte Assessment in der Erfassung von NOS-Konzepten hinsichtlich eines Differential Boosts durch die Testadaptionen?

Für die Frage nach einem möglichen *Differential Boost* wurden sowohl Analysen zur internen Konsistenz als auch zum Mittelwert durchgeführt. Auch wurden Cut-off-Werte definiert. Erfüllten Schülerinnen und Schüler mindestens eines dieser Kriterien wurden sie als Risikolernende definiert. Die internen Konsistenzen wurden ebenfalls über longitudinale Messmodelle bestimmt. Ein *Differential Boost* nach dieser Vorgehensweise liegt genau dann vor, wenn die interne Konsistenz der Risikolernendengruppe zumindest nicht kleiner ist als der Wert der Referenzgruppe. Dies trifft im UDA-Assessment mit Ausnahme der Herkunftsskala zu. Dies gilt allerdings auch für die Entwicklungsskala im Originalassessment. Insbesondere vor dem Hintergrund der negativ formulierten Sicherheitsskala ist dies auffällig und zeigt, dass die Testadaptionen die Hürde der Negativformulierung der Items abbauen kann.

Es zeigt sich aber auch, dass es Zusammenhänge zwischen der Wortanzahl eines Items und dessen Formulierungsrichtung gibt. So scheinen lange Items eher von den Testadaptionen zu profitieren (*Sicherheits-*, *Entwicklungs-* und *Rechtfertigungsskala*), als kurze (*Herkunftsskala*). Während jedoch die Vergleiche zur internen Konsistenz der *Entwicklungs-* und *Rechtfertigungsskalen* zwischen den Risikolernenden und den übrigen Schülerinnen und Schülern ähnliche Werte aufweisen, gilt dies für die *Sicherheits-* und *Herkunftsskala* nicht. Hier weisen die übrigen Lernenden schlechtere Werte auf und es kann eine Benachteiligung durch die Testadaptionen vermutet werden.

Auf Basis der Analyse zu den Mittelwerten zeigt sich, dass sich keine statistisch relevanten Unterschiede zwischen den Gruppen Risikolernende und den Anderen in beiden Assessments beobachten lassen. Vergleicht man darüber hinaus auch die Mittelwerte getrennt nach Gruppen über beide Testversionen, erweisen sich diese jedoch als signifikant unterschiedlich, was sich aber ins Bild der vorherigen Analysen fügt.

Eine letzte Möglichkeit bestand darin, über die latenten Wachstumsmodelle einen *Differential Boost* beobachten zu können. Wenn die Kovarianzen signifikant waren, dann allerdings positiv und nicht negativ. Entsprechend kann auf Basis dieser Modelle kein *Differential Boost* beobachtet werden.

Forschungsfrage 8: Inwiefern unterscheiden sich die UDL-Lernumgebung und die MR-Lernumgebung in der interindividuellen Elaborierung von NOS-Konzepten unter Einbezug von ausgewählten Lernendenmerkmalen?

Die interindividuellen änderungen wurde über Panelmodell unter Kontrolle der *Lesefähigkeit*, der *Intelligenz*, des *sozioökonomischen Hintergrunds* und des *sonderpädagogischen Unterstützungsbedarfs* geschätzt. Zwar besteht ein geringer Einfluss der *Lesefähigkeit* und der *Intelligenz* in einigen Modellen. Den größten Einfluss auf den ersten Messzeitpunkt jedoch hat der *sonderpädagogische Unterstützungsbedarf* bezüglich der *Sicherheitsskala* im UDA-Assessment und der *Entwicklungsskala* im Originalassessment; in beiden Fällen jedoch zu Gunsten der Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf. Den größeten Einfluss auf den zweiten Messzeitpunkt hat in allen Modellen der erste Messzeitpunkt. Außerdem erzielen Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf auf der *Rechtfertigungs-* und *Sicherheitsskala* im Originalassessment signifikant schlechtere Werte. Im UDA-Assessment ist dies nicht der Fall. Die *Art der Lernumgebung* hat keinen sinifkanten Einfluss auf die interindividuelle Elaborierung der NOS-Dimensionen.

Forschungsfrage 9: Inwiefern unterscheiden sich die UDL-Lernumgebung und die MR-Lernumgebung in der intraindividuellen Elaborierung von NOS-Konzepten unter Einbezug von ausgewählten Lernendenmerkmalen?

Für den intraindividuellen Lernerfolg wurden latente Wachstumsmodelle geschätzt. Bezüglich eines *Lehrens und Lernens für alle* zeigen alle Skalen positive intraindividuelle Entwicklung, die jedoch im UDA-Assessment ausgeprägter sind. Demnach lassen sich für beide Assessmentversionen ca. zweidrittel der Schülerinnen und Schüler im Originalassessment bestimmen, die nach der Arbeit mit einem der E-Books höhere Testwerte erzielen. Im UDA-Assessment sind es hingegen mindestens 75 %. Während mit Ausnahme der Herkunftsskala kein Lernender im UDA-Assessment zum zweiten Messzeitpunkt einen schlechteren Wert erzielt, trifft dies im Originalassessment auf jede Skala zu.

In einem weiteren Schritt wurde der Einfluss der Wahrnehmung des *persönlichen Nutzens des E-Books* sowie der *kognitiven Aktivierung* und die *Art der Lernumgebung* auf die NOS-Konstrukte bestimmt. Die intraindividuellen änderungen werden hauptsächlich durch die *kognitive Aktivierung* erklärt. Je höher die Wahrnehmung dieser ist, desto höher ist auch der erzielte Testwert zum zweiten Messzeitpunkt. Der Einfluss der Lernumgebung zum zweiten Messzeitpunkt ist nur bezüglich der *Rechtfertigungsskala* statistisch relevant. Hier zeigen Schülerinnen und Schüler, die mit der *UDL-Lernumgebung* arbeiteten, eine geringere änderungsrate aber keinen geringeren longitudinalen Mittelwert.

Insgesamt spielt jedoch die Art der Lernumgebung keine statistisch bedeutsame Rolle für die intraindividuelle Elaborierung von NOS-Konzepten. Allerdings findet auch keine fokussierte Förderung der *Rechtfertigungsdimension* statt.

6 | Diskussion & Ausblick

Zusammenfassung

Der Ausgangspunkt für dieses Projekt bestand darin, dass es in Deutschland in den letzten Jahren viele politische Entscheidungen und gesellschaftliche Prozesse gab, die immer wieder zu neuen Lerngruppen mit einer Vielzahl von Diversitätsmerkmalen der Schülerinnen und Schüler geführt haben. Hierzu zählen unter anderem die Einführung neuer Schulformen, die Umsetzung der UN-Behindertenrechtskonvention (Inklusion) oder die Auswirkungen der Flüchtlingsbewegungen (Migration). Gleichzeitig sind in Deutschland immer noch die Lernleistungen der Schülerinnen und Schüler z.B.: von dem sozioökonomischen Status (SES), der Lesefähigkeit oder dem Migrationshintergrund anhängig (Autorengruppe Bildungsberichterstattung, 2016, 2018). Vor dem Hintergrund einer Bildungsgerechtigkeit und ausgehend von einem positiven Menschenbild, das Unterschiede als wertvoll betrachtet, ist dieser Umstand nicht akzeptierbar.

Aus dieser Motivation heraus zielt dieses Projekt darauf, das Lernen und die Messung von Lernleistungen neu zu konzeptualisieren und anschließend zu evaluieren. Obwohl diese Problematik schon länger besteht, sind bisher eher theoretische Konzeptionen für einen *diversitätswertschätzenden Unterricht* als Unterrichtsmodell formuliert worden (Fischer et al., 2014; Sliwka, 2012). Außerdem bleibt oftmals die Verbindung eines solchen Unterrichts zu den Merkmalen von *Gutem Unterricht* offen (Klieme & Rakoczy, 2008) (Tab. 2.1).

Eine vielversprechende Möglichkeit stellen hierzu das Universal Design for Learning (UDL) zur Erstellung von Lernumgebungen und Universal Design for Assessment (UDA) zur Erstellung von Assessments dar. Sowohl UDL als auch UDA setzen konzeptionell beim Gedanken der Barrierefreiheit an. Damit sind jedoch nicht nur physische Barrieren gemeint, sondern auch solche, die Lernumgebungen von sich aus umfassen, wenn sie im One-size-fits-all-Ansatz erstellt wurden. Während UDL durch die Prinzipien der *multiplen Repräsentation von Informationen* ("Was" des Lernens), der *multiplen Wege der Verarbeitung von Informationen und der Darstellung von Lernergebnissen* ("Wie" des Lernens) sowie der *multiplen Wege zur Förderung des Lernengagements und der Lernmotivation* ("Warum" des Lernens) (Tab. 2.2) (CAST, 2018; Schlüter et al., 2016) Barrieren abbaut, fokussiert UDA auf die Reduzierung der *konstruktirrelevanten Varianz* (KIV) und bietet somit einen theoretischen Rahmen, um die *Testzugänglichkeit* zu gewährleisten (Tab. 2.3) (Beddow, 2011; Lovett & Lewandowski, 2015; Thompson et al., 2004).

UDL nimmt bezüglich der Diversität von Lerngruppen eine *positive Perspektive* ein. Diversitätsmerkmale werden als Ressource für den Unterricht genutzt. Entsprechend sind nicht die individuellen Unterschiede für Wohl und Wehe des Unterrichts verantwortlich, sondern der Unterricht muss diese adäquat und multidimensional aufgreifen (Hall et al., 2012). In gleicher Weise fordert UDA die Entwicklung von zugänglichen Items, Skalen

und Assessments. Daher sollten Items wenig oder keinen Inhalt enthalten, der unnötig ist, um Fähigkeiten, Wissen oder Einstellungen zum Zielkonstrukt zu demonstrieren. Dies ist besonders wichtig in Fällen, in denen Zugangskompetenzen für den Prüfling eine Herausforderung darstellen. Ein typisches Beispiel ist die Notwendigkeit, einen erzählenden Text zu lesen, um eigentlich ein mathematisches Problem zu lösen (Elliott et al., 2018). Im Beispiel stellt demnach die Lesefähigkeit eine konstruktirrelevante Fähigkeit für die Testaufgabe dar; der Test misst ein mathematisches Konstrukt.

Um UDL-basierte Lernumgebungen und UDA-basierte Assessments zu konzeptualisieren, wurde der Inhaltsbereich Nature of Science (NOS) ausgewählt. NOS spielt einerseits in vielen curricularen Vorgaben eine zentrale Rolle als Element der naturwissenschaftlichen Grundbildung (Abd-El-Khalick & Lederman, 2000; Bernholt et al., 2012; Holbrook & Rannikmae, 2007; Lederman, 2013; NGSS Lead States, 2013). Andererseits wurde NOS, obwohl es für die fachdidaktische Konzeption der naturwissenschaftlichen Fächer eine zentrale Rolle einnimmt, bisher nicht in Interventionen kombiniert mit UDL eingesetzt. Für die Gestaltung der Lernumgebung und des Assessments wurden die NOS-Inhaltsbereiche nach Kampa et al. (2016) verwendet, die die *Sicherheit des naturwissenschaftlichen Wissens*, die *Entwicklung von naturwissenschaftlichem Wissen*, die *Rechtfertigung des naturwissenschaftlichen Wissens* und die *Herkunft des naturwissenschaftlichen Wissens* umfassen. Mit Blick auf die naturwissenschaftliche Grundbildung spielt die Dimension der *Rechtfertigung des naturwissenschaftlichen Wissens* eine besondere Rolle.

Da sich aus dem Forschungsfeld zu UDL eine mögliche Abhängigkeit zwischen der Intervention und dem dazugehörigen Assessment ableiten lässt und sich darüber hinaus ein Fokus auf dem UDL-Prinzip der *multiplen Repräsentationsformen* beobachten lässt (Al-Azawei et al., 2016; Capp, 2017; Rao et al., 2014), wurde ein 2x2-Between-Subject-Design gewählt. Hierüber lässt sich die mögliche Abhängigkeit zwischen Lernumgebungen und adaptierten und nicht-adaptierten Assessments bestimmen (UDA-Assessment und Originalassessment). Außerdem lässt die Frage nach einem hypothetischen Vorteil eine ganzheitliche UDL-Lernumgebung gegenüber einer solchen untersuchen, die durch die Verwendung eines Videos auf das Prinzip der *multiplen Repräsentationen* (MR-Lernumgebung) fokussiert (King-Sears et al., 2015).

Für die Analyse der Daten wurden Methoden gewählt, die einerseits einen möglichst exakten Vergleich zwischen den vier Gruppen zu beiden Messzeitpunkten ermöglichen. So wurden die instruktional sensitiven Items bestimmt (Deutscher & Winther, 2018; Naumann et al., 2017; Polikoff, 2010). Auf Basis dieser Analysen wurden in einem zweiten Schritt die Skalen neuformuliert. Über konfirmatorische Faktorenanalysen (KFA) im *Multiple Indicator, Multiple Cause*-(MIMIC)- bzw. im *Multiple Group*-(MG)-Ansatz (Brown, 2015; Kline, 2016) und in der Kombination ist dies trotz einer eher kleinen Stichprobengröße möglich (Sideridis et al., 2015). Um die Gruppenabhängigkeit der Items und damit die *Testzugänglichkeit* zu bestimmen, wurden ausgehend von den erhobenen Lernendenmerkmalen Subgruppen gebildet. Die Messinvarianz (MI) der Items wurde über ein Partial-Credit-Modell grafisch bestimmt (Schwab & Helm, 2015). Damit aber nicht nur die interindividuellen Entwicklungen Beachtung in den Analysen finden, wurden auch latente Wachstumsmodelle eingesetzt, um die intraindividuelle Entwicklung zu bestimmen und ebenfalls vor dem Hintergrund von Lernendenmerkmalen zu kontrollieren (Wu et al., 2010).

Vor dem geschilderten Hintergrund kann zusammenfassend festgehalten werden, dass

sich die Lernumgebungen in der Hauptstudie nicht in der Förderung von NOS-Konzepten unterscheiden. Entsprechend besteht kein Vorteil der UDL- gegenüber der MR-Lernumgebung. Hierfür sind zwei Erklärungen möglich. Erstens bestand schon in der Vorstudie ein Vorteil der videobasierten Inhaltsform gegenüber den anderen. Außerdem umfassten beide Lernumgebungen das gleiche Video und die UDL-Lernumgebung darüber hinaus mehr Funktionen, daher kann eine mögliche Begründung in dem wahrgenommenen Nutzen selbst liegen. Dies deutete sich auch bereits in der Vorstudie an. Hier unterscheiden sich die Lernumgebungen nicht signifikant voneinander. Eine mögliche Schlussfolgerung ist, dass der Funktionsumfang nicht genutzt wurde bzw. explizit eingeführt und in seiner Verwendung ritualisiert werden muss. Wenn das Video das Hauptmerkmal der Lernumgebung ist, dann vermag es andere Effekte zu überlagern. Ein Vergleich mit einer Lernumgebung ohne Video – zum Beispiel unter Verwendung des Comics oder des Pop-Up-Textes als Inhaltsform – hätte unter Umständen an dieser Stelle größere Unterschiede hervorgebracht. Positiv formuliert, führt jedoch der größere Funktionsumfang auch nicht zu einem schlechteren Abschneiden. Weiterhin werden auch nicht ausschließlich Konzepte zum Bereich der *Rechtfertigung des naturwissenschaftlichen Wissens* gefördert. Die intendierte, explizite Förderung von Rechtfertigungskonzepten blieb aus. Möglicherweise ist die Darstellung nicht ausreichend explizit.

Die Likertskalen zur Erfassung der NOS-Dimensionen werden im Rahmen dieser Arbeit im Sinne einer Quasi-Leistungsmessung verwendet. Die Items der Skalen sind nach den Autoren als wissensbasierte Vorstellungen veranlagt (Urhahne et al., 2008). Entsprechend weisen sie einerseits einen *affektiven* Charakter auf. Exemplarisch sei die *Herkunftskala* genannt, die explizit auch Anfänger als Produzenten für naturwissenschaftliches Wissen benennt. Entsprechend muss auch der/die Proband/in abschätzen, ob er oder sie bereit ist, eine naturwissenschaftliche Frage- oder Problemstellung zu untersuchen. Der andrerseits *kognitive* Charakter der Skalen wird exemplarisch an der *Entwicklungsskala* deutlich, in der die/der Proband/in ihr/sein eigenes Wissen mit den Beobachtungen und den daraus gezogenen Schlüssen beurteilen muss.

Außerdem bestand die Frage nach der Vergleichbarkeit der Testinstrumente. Diese wird durch die Daten der latenten Konstrukte unterstützt. So sind trotz diverser Adaptionen und der Portierung eines analogen Tests auf ein digitales Format die latenten Konstrukte in den Vollskalen hinsichtlich der Faktorladung und deren Struktur vergleichbar. Allerdings zeigt sich auch, dass das Verhältnis von der Varianz der Stichprobe und der Anzahl der Items sich verschlechtert. Durch die Bestimmung der instruktional sensitiven Items kann dennoch mit dem Problem der zu großen Varianz umgegangen werden. Nach der Bestimmung der instruktional sensitiven Items und der Formulierung von Kurzskalen mit je vier Items sind darüber hinaus auch die Nullpunkte und die Fehlervarianzen zumindest partial messinvariant. Die Bestimmung der instruktionalen Sensitivität der Items stellt entsprechend einen wesentlichen Schritt in der Auswertung dieser und anderer (UDL-)Interventionen dar (Al-Azawei et al., 2016; Capp, 2017; Rao et al., 2014).

Betrachtet man die Mittelwertsunterschiede auf latenter Ebene so unterscheiden sich die Assessmentversionen einerseits in den Differenzen zwischen beiden Messzeitpunkten als auch in der Varianz und damit in der Effektstärke (Tab. 5.37 und 5.38). Da sich die Lernumgebungen im *MIMIC*-Ansatz als messinvariant erwiesen haben, wurde die Analyse auf die Assessmentversionen reduziert. Im UDA-Assessment liegen jeweils kleine Effekte vor, während im Originalassessment dies nur für die *Entwicklungsskala* gilt. Dies kann als Folge der besseren internen Konsistenzen verstanden werden. Größere Effekte sind durch

die Kürze der Intervention nicht erwartbar gewesen.

Der Einfluss der erhobenen Lernendenmerkmale auf die inter- wie die intraindividuelle Elaborierung ist gering. Lediglich die *wahrgenommene kognitive Aktivierung* zeigt einen wesentlichen Einfluss auf die Effekte in den latenten Wachstumsmodellen zum Originalassessment; allerdings kann nicht sicher gesagt werden, ob sich die erhöhten Anstrengungen auf das Assessment beziehen oder auf das E-Book. Dies begründet sich darin, dass die Items dieser Skala explizit die E-Books adressieren. Weil jedoch iPads sowohl für die Testung als auch für die Lernumgebung verwendet wurden, kann es sein, dass die Schülerinnen und Schüler beides zusammen eingeschätzt haben. Diese Annahme würde erklären, warum die *wahrgenommene kognitive Aktivierung* sich signifikkant nach Assessmentversion, aber nicht nach Lernumgebung unterscheidet. Im Originalassessment weisen wenige Items - vor allem vor dem Hintergrund des *sozioöknomischen Statuses* - eine Gruppenabhängigkeit auf. Allerdings zeigt sich ein Schereneffekt beim Originalsssessment bezüglich der intraindividuellen Entwicklung, welcher sich im UDA-Assessment nicht zeigt. Da dieser Schereneffekt auch eine Verschlechterung einiger Schülerinnen und Schüler aufzeigt (Tab. 5.39), ist er in dieser Form nicht gewünscht (Abb. 5.14).

Hierfür sind jedoch folgende Erklärungen denkbar. Zuerst steigt die Varianz im Orginalassessment vom ersten zum zweiten Messzeitpunkt. Dies wiederum könnte einerseits damit begründet werden, dass die Lernenden zum zweiten Messzeitpunkt weniger motiviert gewesen sind und deshalb extremer geantwortet haben. Andererseits kann vermutet werden, dass beide Lernumgebungen erst ein Verständnis dafür geschaffen haben, was die Items und damit die Skalen im Orginalassessment aussagen. Beide Fälle sind aus einer psychometrischen Perspektive problematisch, weil einerseits ohne Vorerfahrung oder -wissen das Konstrukt nicht ausreichend gemessen werden kann und andererseits das Testformat demotivierend wirken kann (Baumert & Demmrich, 2001). Aber auch aus einer didaktischen Perspektive weist dieser Befund Aspekte auf. So findet über die Lernumgebungen eine Wissensentwicklung statt (Lernsituation), jedoch wird auch die Bedeutung der Leistungssituation betont. Aus einer diversitätswertschätzenden Perspektive heraus, muss dies ebenfalls bedacht werden.

Das Vorhandensein äquivalenter Konstrukte im UDA- und Originalassessment ist eine notwendige Voraussetzung für den Vergleich der Daten. Nur angenommene Gleichheit ist ohne empirische Bestätigung nicht ausreichend (Dorans & Middleton, 2012). Es reicht daher nicht aus, nur die *konfigurale MI* zu akzeptieren (Lovett & Lewandowski, 2015). Um die Daten vergleichbar zu machen, muss zumindest eine *partial-skalare MI* vorhanden sein, damit die Mittelwerte interpretierbar sind. Vor dem Hintergrund des Forschungsfelds um NOS, zeigt sich, dass Likertskalen geeignet sind, die Effekte aus der Intervention aufzuzeigen. Der Grad des Effektes hängt jedoch mit der *Art des Instruments* und nicht mit der *Art der Lernumgebung* zusammen.

Obwohl zwischen den Forschungsfeldern von UDL und NOS Gemeinsamkeiten identifiziert werden konnten, wurde die eigentlich adressierte NOS-Dimension *Rechtfertigung* nicht ausschließlich gefördert. Durch die Intervention erfolgte eine Förderung aller NOS-Dimensionen. Es kann daher einerseits geschlussfolgert werden, dass eine empirische und theoretische Verbundenheit der Dimensionen vorliegt und eine gezielte Förderung von NOS-Inhaltsbereichen verhindert (Harrison et al., 2015). Andererseits kann aber auch vermutet werden, dass die Elemente der Lernumgebungen die übrigen Inhaltsbereiche streifen und nicht ausreichend fokussiert sind. Auch ein möglicher Re-Testeffekt durch die

kurz aufeinanderfolgenden Tests kann nicht gänzlich ausgeschlossen werden, wobei diese Vermutung eher auf das Originalassessment zutrifft; die Änderungen im UDA-Assessment sind hierfür zu gleichförmig.

Limitierungen

Auch diese Studie weist wesentliche Limitierungen auf. Vor dem Hintergrund von *UDL* und *NOS* bleibt die Frage nach der Nachhaltigkeit der Intervention offen. Dieser Verdacht betrifft jedoch verstärkt das *Originalassessment*. Aus technischen Gründen war es auch nicht möglich, Informationen zur konkreten Art der Bedienung der digitalen Lernumgebungen zu sammeln. Dies resultiert aus den technischen Limitierungen der verwendeten App. Beispielsweise hätte eine Webseite als Lernumgebung den Einsatz von Cookies ermöglicht. Alternativ wäre in einer Folgestudie der Einsatz von Videos möglich, um die Nutzung der Ebooks beschreiben zu können.

Weiterhin besteht keine Vergleichsmöglichkeit zu analogen Lernformaten. Zwar wurde auf diesen Vergleich aus guten Gründen verzichtet. Hierzu wäre eine größere Stichprobe notwendig gewesen. Außerdem hätte die Intervention wohlmöglich unter Differenzen bei der Motivation der Probanden gelitten, wenn Lernende mit und ohne iPad zusammengearbeitet hätten. Oder aber der Randomisierungsgrad der Gruppen in der Stichprobe wäre geringer, um das letztgenannte Problem zu umgehen. Jedoch wird das papierbasierte Arbeitsblatt vermutlich noch eine Weile die dominierende Form bleiben. Es bleibt daher die Frage offen, ob die technischen Funktionen der Lernumgebungen oder aber die fachdidaktische Veranlagung die beobachteten Effekte verursacht haben.

Schließlich wurde bisweilen nur die Inputphase eines UDL-basierten Unterrichts überprüft. Dies geschah aus dem Grund, weil angenommen wurde, dass die Zugänglichkeit zu den Lerninhalten ein entscheidender Schritt im Unterrichtsgang ist. Darüber hinaus sollte jedoch noch die Outputphase untersucht werden, in der die Schülerinnen und Schüler Lernprodukte erstellen. Dies böte die Möglichkeit, über eine Analyse der Artefakte weitere Daten insbesondere zur Nachhaltigkeit und zu möglichen Unterschieden der MR- und der UDL-basierten Lernumgebung zu gewinnen.

Ebenso wie bei der Ebook-App war es auch bei der Assessmentapp nicht möglich, den Bearbeitungsprozess zu tracken. Außerdem liegen keine Daten zu den metakognitiven Prozessen der Testteilnehmer bei der Durchführung des *UDA*- und des *Originalassessments* vor. Diese könnten zum einen mit Hilfe von Think-Aloud-Studien oder zum anderen durch die Testaufgaben für die evozierte kognitive Belastung erhoben werden. In beiden Fällen besteht jedoch das herausfordernde Problem, dass streng zwischen der gewünschten Interaktion des Testteilnehmers mit dem Zielkonstrukt und den unerwünschten Nebenlasten unterschieden werden muss. Darüber hinaus liegen keine Daten darüber vor, welche der Maßnahmen aus der Testadaption einen signifikanten Anteil der Auswirkungen hat. Oder auch welche Testanpassungen verwendet wurden und wie oft.

Außerdem kann die Wahl der Diversitätsdimensionen als nicht unzureichend angesehen werden. Wichtige Dimensionen könnten fehlen, die hier nicht aufgeführt sind. Schließlich kann auch argumentiert werden, dass eine Bewertung auf iPads keine vergleichbaren Ergebnisse zu früheren Studien liefert, da die Motivation der Testteilnehmer allein schon durch den Einsatz von den Geräten steigt. Dieses Problem ist ernst in dem Sinne, weil kein Vergleich zu einer Papier-Stift-Version der Assessments möglich ist. Ein potenzieller Motivationseffekt müsste jedoch bei allen Testteilnehmern durch den Einsatz von

iPads vorliegen. Schließlich stellt sich die Frage der Übertagbarkeit dieser Ergebnisse auf Leistungstests; hier wird oftmals von einem größeren Einfluss der Hintergrundvariablen berichtet (Nehring et al., 2015)

Während diese Studie lediglich Modelle für die einzelnen NOS-Dimensionen vor dem Hintergrund der Lernendenmerkmale schätzen konnte, können darüber hinaus keine Aussagen zu den Effekten der Lernendenmerkmale in einem Globalmodell mit allen NOS-Dimensionen getroffen werden. Hier wären unter Umständen noch weitere Detailbeobachtungen möglich. Dafür ist jedoch die Stichprobengröße nicht ausreichend, bzw. die Datenlage aus beiden Assessments zu unterschiedlich. Dies hat sich bereits bei der Abbildung der NOS-Modelle mit allen Items gezeigt. Zwar wurden diese theoriegeleitet gekürzt, jedoch führte diese Kürzung auch zu einem Verlust an Informationen über die Stichprobe. Schließlich lassen Daten und Methoden auch keine Aussagen dazu zu, in welcher Art und Weise sich die Förderung der übrigen NOS-Dimensionen über die Adressierung der Rechtfertigungsdimension erfolgte. Hierzu müssten Interviews erfolgen, in denen die Probanden ihre subjektive Sicht auf die Lernumgebung und deren Verbindung zu den NOS-Dimensionen beschreiben. Letztlich bleibt die grundsätzliche Frage, warum Schülerinnen und Schüler in kontextualisierten Situationen etwas über NOS lernen sollten, um daraufhin dekontextualisiert befragt zu werden. Zukünftige Testentwicklungen sollten diesen Punkt stärker in Betracht ziehen, gleichzeitig aber Testkonstruktionen vermeiden, die genuin Fächer oder NOS-Wissen zu Kontexten prüfen.

Schlussfolgerungen

Sowohl in der Vor- als auch in der Hauptstudie konnte gezeigt werden, dass die UDL-basierten Lernumgebungen unabhängig vom Assessment zu einer Elaboration der NOS-Konzepte führen. Im Sinne der Kriterien eines diversitätsschätzenden Unterrichts (Tab. 2.1) besteht in einem UDL-basierten Unterricht die Möglichkeit die Lernenden als KonstrukteurInnen ihres Wissens einzusetzen. Außerdem kann diese Konstruktion sozial eingebetet werden. Und schließlich besteht durch die herausfordernde Gestaltung eines solchen Unterrichts einerseits eine gewisse Planbarkeit und damit andererseits Sicherheit im unterrichtlichen Handeln.

Im Rahmen dieser Studie hat sich die Dimension der *kognitiven Aktivierung* als relevant erwiesen. Diese gilt als eine der Basisdimensionen für qualitätsvollen Unterricht (Klieme & Rakoczy, 2008; Praetorius & Charalambous, 2018). Im Sinne eines *Lernens und Lehrens für alle* spielt die *kognitive Aktivierung* eine mögliche Schlüsselrolle. In dieser Studie zeigt sich, dass Schülerinnen und Schüler mit einer hohen wahrgenommenen kognitiven Aktivierung auch eine höhere Testleitung im Prä-Post-Vergleich im Originalassessment zeigen.

Neben der Gestaltung von Unterricht im Sinne einer *Inclusive Education* unter Hinzunahmen des Gedankens von Barrierefreiheit, spielt auch die Leistungsmessung eine Rolle. Das UDA-Assessment weist bessere interne Konsistenzen als auch bessere Faktorladungen auf. Damit werden sowohl auf manifester als auch auf latenter Ebene bessere Gütekriterien für die Reliabilität und Validität erreicht. Dies begründet sich in der von Beginn an fokussierten Zugänglichkeit des Assessments. Während für Studien, die in von Diversität geprägten Lerngruppen stattfinden, dieser Gedanke und die sich daraus ergebenen Konsequenzen eine herausragenden Rolle spielen, können wohlmöglich viele andere Studien auch von diesem profitieren.

Abschließend sei noch die Frage gestellt, welchem der Assessments der Vorzug gegeben werden sollte bzw. welches sich als valider und realibler erweist. Das UDA-Assessment wurde mit Blick auf die Diversität der Testeilnehmer konstruiert. Die inhaltliche Genauigkeit wurde durch Expertenratings sichergestellt. Dieser Grundgedanke äußert sich in der Konzeption durch bessere Lesbarkeit (Textformatierungen), durch alltagsnähere Sprache (Verständlichkeit) und einem zumindest hypothetisch bekannteren Antwortformat. Das Originalassessment wurde hingegen ausschließlich vor dem Grundgedanken einer möglichst genauen inhatlichen Erfassung der NOS-Konzepte konstruiert.

Die Stichprobe erweist sich als divers. Dies gilt im Besonderen für die Daten zur Lesefähigkeit und der Intelligenz der Probanden. Entsprechend ist das UDA-Assessment, was die Gedanken zur Gewährung von Testzugänglichkeit und zur Vermeidung von konstrukttirrelevanter Varianz von Anfang an miteinbezogen hat, bevorteilt. Dies äußert sich rein optisch in einer besseren Übersichtlichkeit und damit Lesbarkeit des UDA-Assessments und statistisch in den besseren internen Konsistenzen (mit Ausnahme der Herkunftsskala). Daraus resultierend ergibt sich eine genauere Beschreibung der NOS-Konzeptentwicklung.

Zusammengefasst kann mit *UDA* die Zugänglichkeit von Tests erhöht werden, ohne das Zielkonstrukt zu ändern. Dies ist eine wichtige Voraussetzung, um unsystematische Bewertungsmuster von Items zu vermeiden (Lamprianou & Boyle, 2004). Dies umfasst sowohl die psychometrischen Merkmale von Tests als auch die tatsächliche Erfassung des Konstrukts. Weniger komplexe Adaptionen wie die Verlängerung der Bearbeitungszeit oder Reduzierung der Itemanzahl scheinen dabei weniger geeignet zu sein (Anderson, Lai, Alonzo & Tindal, 2011; Bridgeman, Trapani & Curley, 2004; Wise & Kingsbury, 2016). Bei der Entwicklung von Leistungstests könnte daher auch die Verwendung alternativer Darstellungen für Texte oder von Leichter Sprache sowie die Implementierung von Vorlesefunktionen von Anfang an in Betracht gezogen werden. Die Formulierung von Likertitems sollte nicht negativ sein (Salas-Wright et al., 2013). Die Skalen zur Sicherheit und zur Herkunft von naturwissenschaftlichem Wissen weisen schlechtere interne Konsistenzen auf - auch weil sie negativ formuliert sind. Während jedoch die Testadaption bei längeren Items den Effekt einer verbesserten internen Konsistenz der Skala aufweist (Sicherheit), gilt dies für die Herkunftsskala nicht. Elemente zur Seitenorganisation sind für eine effektive Unterstützung des/der Testteilnehmenden geeignet. Schließlich können alternative Antwortformate in Betracht gezogen werden, die durch den Einsatz digitaler Mittel ausgewertet werden können. So könnten beispielsweise Techniken aus dem maschinellen Lernen oder Deep Learning verwendet werden (z.B.: mit neuronalen Netzen), um Zeichnungen zu bewerten, wenn diese in größerem Maßstab anfallen.

Welche Rückschlüsse lassen sich auf *UDL* ziehen? Die genutzten Lernumgebungen ermöglichen vor allem eins: Redundanz. Diese findet im Frontalunterricht kaum statt. Hingegen wird im kanonischen Arbeiten der Gleichschritt aller Lernenden verlangt (Menthe & Sander, 2016). Das Lernen in Kleingruppen mit hochstrukturierten Lernmaterialien, wie sie durch UDL zur Verfügung gestellt werden, bietet hierbei das Potenzial Redundanz und Fokussierung bereitzustellen. Die in dieser Studie verwendeten Inhaltsrepräsentationsformen zeigen, dass insbesondere digitale Medien die Möglichkeiten zur Herstellung von Barrierefreiheit bieten. Das Video als eine Form der Inhaltsrepräsentation nimmt hierbei eine besondere Rolle ein, wie der Vergleich der UDL- und der MR-Lernumgebung und beiden Studien zeigt.

Mit Blick auf das Unterrichten kann vermutet werden, dass ähnliche Lernumgebungen

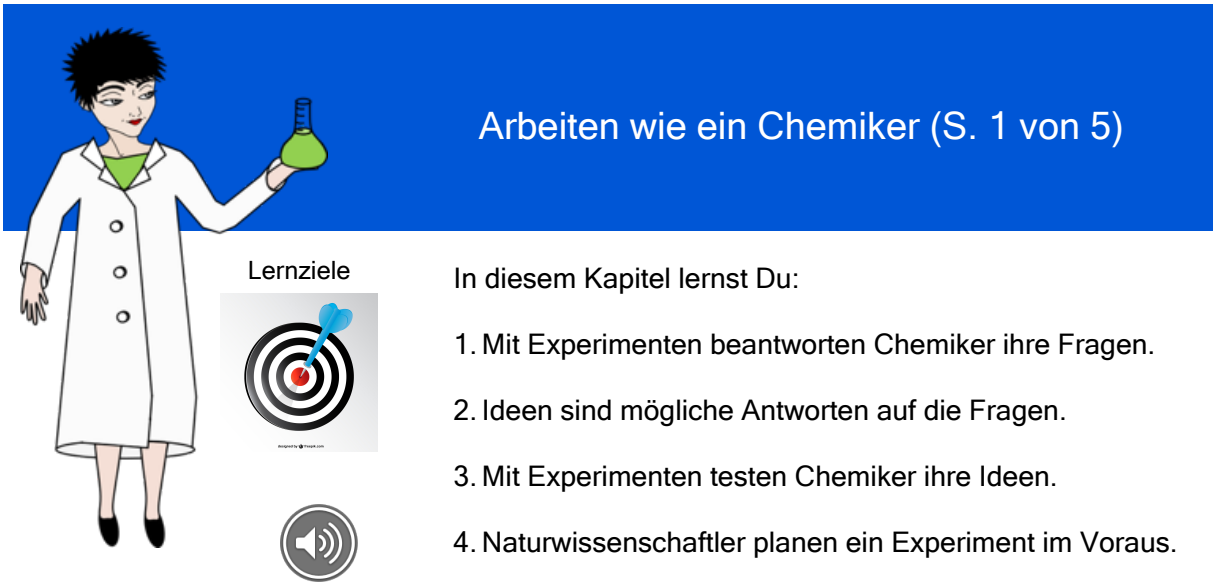
und alternative und selbstgewählte Lernprodukte ein genaueres Bild der Lern- und Leistungsfähigkeit der jeweiligen Schülerinnen und Schüler zulassen. Zwar bieten schriftliche Leistungsüberprüfungen für die Lehrkraft ökonomische Vorteile, jedoch kann Wissen auch über die Erstellung von Artefakten, wie Laborjournalen oder Portfolios effizient demonstriert werden. Hierüber gewinnt die Lehrkraft Informationen zu den Eigenschaften der Schülerinnen und Schüler. Gleichzeitig steigt jedoch der Aufwand in der Bewertung. Jedoch steigt wohlmöglich auch die Motivation und die Einstellung zum Unterricht. Damit wäre sowohl ein diversitätschätzender Unterricht angelegt als auch die Lehrkraft entlastet.

Für die *Elaborierung von NOS-Konzepten* kann aus der Studie geschlossen werden, dass diese einer langwierigen Entwicklung bedürfen. Die in der Studie beobachteten Effekte sind als klein zu bezeichnen. Während dekontextualisierte Tätigkeiten sicherlich geeignete Ansätze darstellen, um in NOS-Problematiken einzuführen, eignet sich die Verwendung von kontextualisierten Tätigkeiten besser, um die Kriterien eines (diversitätswertschätzenden) Unterrichts zu erfüllen, weil diese eine Anschlussmöglichkeit an die Lebenswelt der Schülerinnen und Schüler bieten, was die Initiative der Lernenden und die Bedeutsamkeit des Lerninhalts steigert (Lee et al., 2015). In diesem Fall müssen NOS-Konzepte explizit reflektierend in den Vordergrund der unterrichtlichen Tätigkeit gestellt werden. Vor diesem Hintergrund scheint es lohnenswert, einen zeitlichen Verlauf der NOS-Konzeptentwicklung über die Sekundarstufe zu beschreiben und zu erheben. Bislang fehlen hierzu jedoch Längsschnittsstudien.

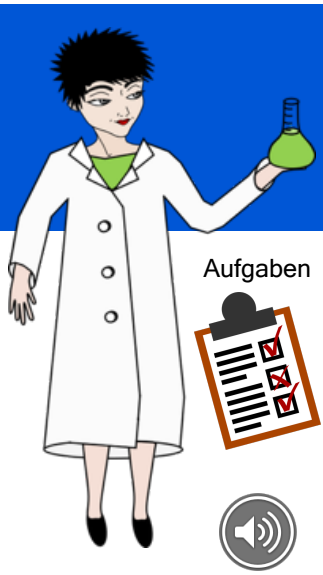
Über das Zusammenspiel von Diversitätsdimensionen, Wissensdomänen oder -fähigkeiten und die Form standardisierter Tests wurde bisher wenig in den Fachdidaktiken der Naturwissenschaften berichtet. Während der sozioökonomische Status, die Intelligenz oder die Lesefähigkeit kaum einen Einfluss auf die NOS-Elaboration in dieser Studie aufweisen, gilt dies für den sonderpädagogischen Unterstützungsbedarf nicht. In der qualitativen, wie in der quantitativen Studie profitierten Schülerinnen und Schüler mit einem solchen Status von den Lernumgebungen.

7 | Appendix

7.1 UDL-Lernumgebung



1



Arbeiten wie ein Chemiker (S. 2 von 5)

Aufgaben

1. Du bist in einer Gruppe von Chemikern.

   Plane ein Experiment mit den Chemikalien und Gerä-

   ten auf dem Tisch.

2. Deine Frage lautet:

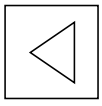
   Ist gleich viel auch gleich schwer?

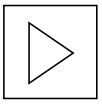
   Testet eine der Ideen von den Chemikern.

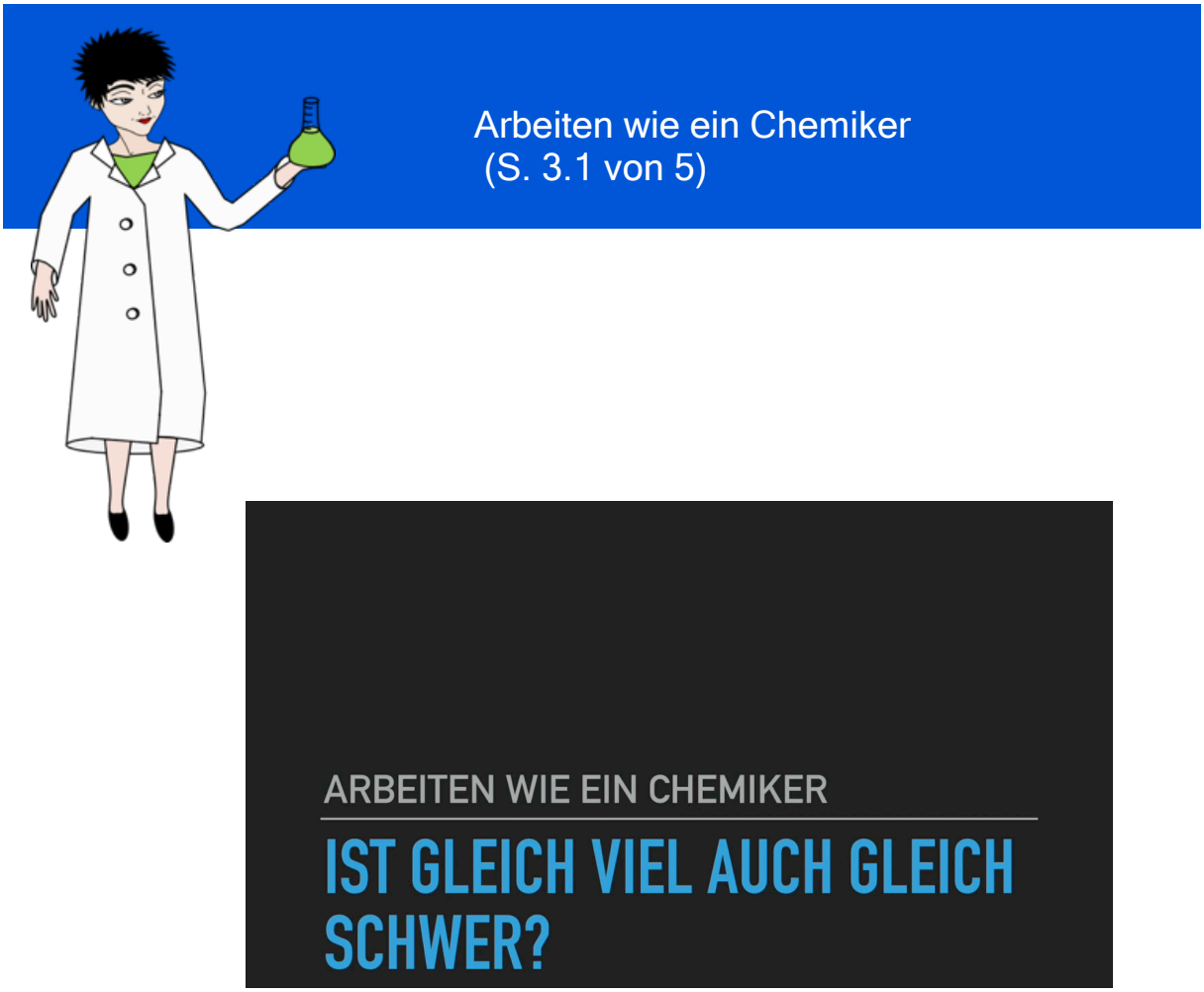
3. Beantwortet mit der Checkliste die Frage:

   Welcher Chemiker hat Recht?

4. Dir helfen ein Video, ein Comic oder ein Text.



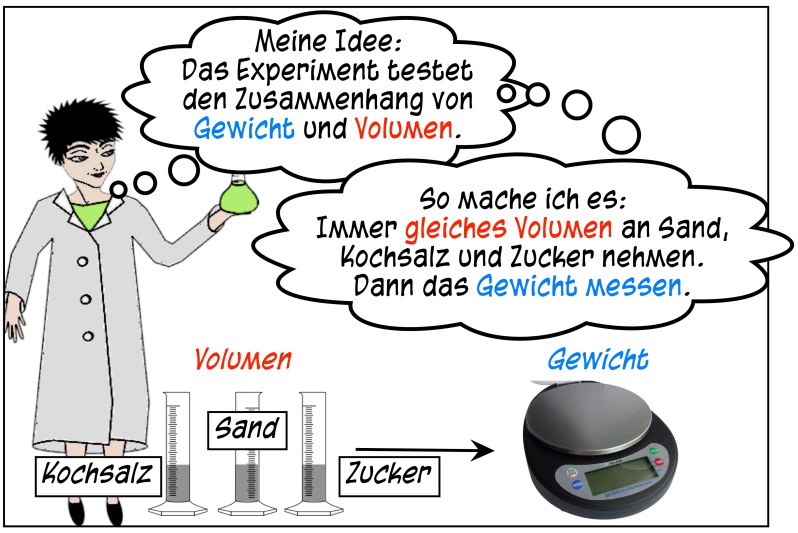




Hier geht es zurück zur Aufgabe.

Hier geht es zum Experiment.

Arbeiten wie ein Chemiker
(S. 3.2 von 5)



Hier geht es zurück zur Aufgabe.

Hier geht es zum Experiment.

Arbeiten wie ein Chemiker
(S. 3.3 von 5)

Wie nutzen Chemiker Experimente?

Klick für Hilfe auf das Fragezeichen!

Wie planen Chemiker Experimente?

Klick für Hilfe auf das Fragezeichen!

Wie misst ein Chemiker Volumen?

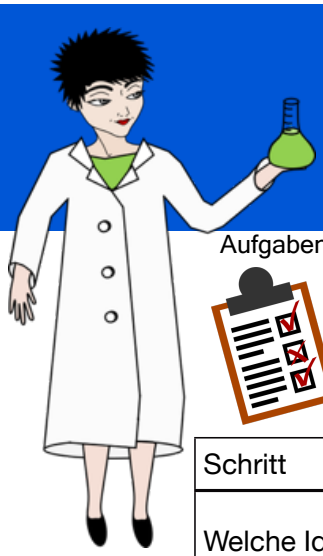
Klick für Hilfe auf das Fragezeichen!

Wie misst ein Chemiker Gewicht?

Klick für Hilfe auf das Fragezeichen!

Hier geht es zurück zur Aufgabe.

Hier geht es zum Experiment.

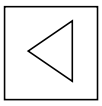


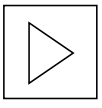
Arbeiten wie ein Chemiker (S. 4 von 5)

Aufgaben

Checkliste zum Experiment

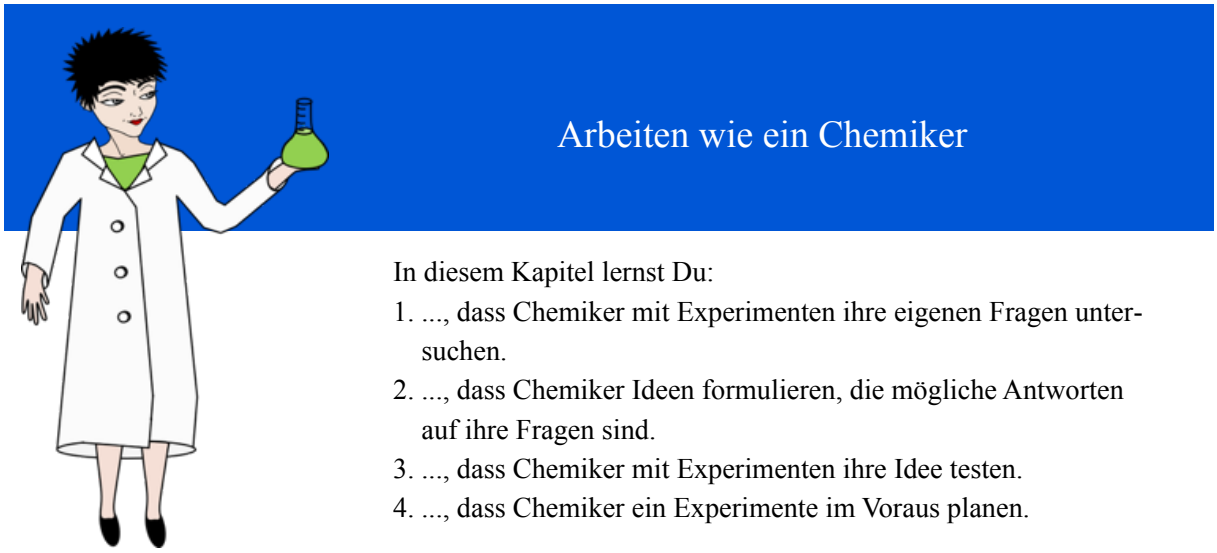
Schritt	Das machen wir
Welche Idee testet ihr?	
Welche Experimentiermethoden nutzt ihr?	
Wie geht ihr vor? Was verändert ihr im Experiment? Was verändert ihr nicht?	
Was könnt ihr beobachten?	
Welcher Chemiker hat recht? Nutzt zur Beantwortung eure Beobachtungen.	





6

7.2 MR-Lernumgebung



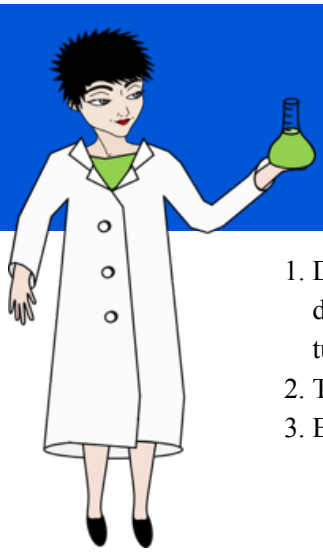
In diesem Kapitel lernst Du:

1. ..., dass Chemiker mit Experimenten ihre eigenen Fragen untersuchen.
2. ..., dass Chemiker Ideen formulieren, die mögliche Antworten auf ihre Fragen sind.
3. ..., dass Chemiker mit Experimenten ihre Idee testen.
4. ..., dass Chemiker ein Experimente im Voraus planen.

---

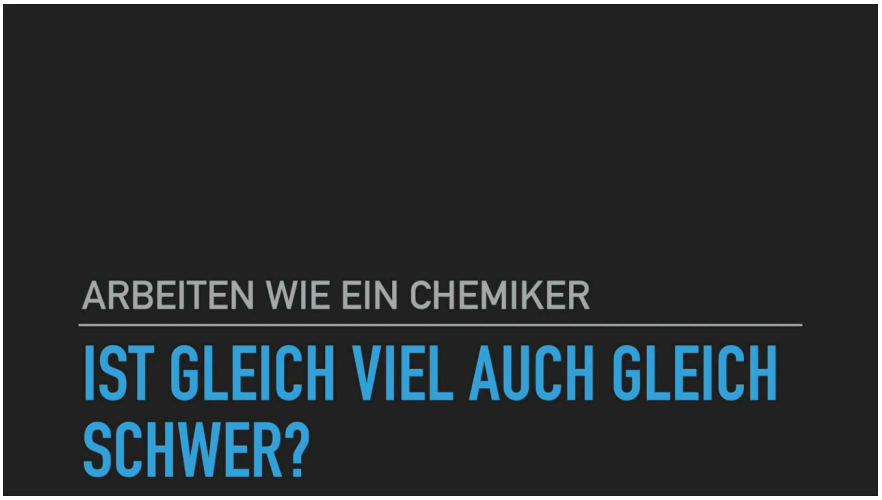
1. Lies den Comic und diskutiere, welche der beiden Ideen auf die Frage, die richtige ist.

1



Arbeiten wie ein Chemiker

1. Du bist Teil einer Gruppe von Chemiker. Plant ein Experiment mit den Chemikalien und Geräten auf dem Tisch, um die Frage der Naturwissenschaftler zu beantworten.
2. Teste eine der beiden Ideen der Chemiker von der vorherigen Seite.
3. Benutze das Video als Hilfe.



Multiple Mittel der Informationsrepräsentation	Multiple Mittel der Verarbeitung von Informationen und der Darstellung von Lernergebnissen.	Multiple Möglichkeiten der Förderung von Lernengagement und Lernmotivation.
1. **Biete Wahlmöglichkeiten bei der Perzeption.** • Biete Möglichkeiten, die Darstellung von Informationen anzupassen. • Biete Alternativen zur auditiven Informationsvermittlung an. • Biete Alternativen zur visuellen Informationsvermittlung an.	4. **Ermögliche unterschiedliche motorische Handlungen.** • Variiere die Möglichkeiten zur Steuerung von Lernmaterialien. • Variiere die Möglichkeiten zur Erstellung von Antworten. • Optimiere den Zugang zu Lernhilfen, Lernmedien und technischen Hilfsmitteln (angepasste Tastaturen etc.).	7. **Biete variable Angebote zum Wecken von Lerninteresse.** • Eröffne möglichst viele Wahlmöglichkeiten und räume möglichst viel Autonomie ein. • Biete möglichst relevante, positiv bewertete und authentische Aufgaben und Aktivitäten an. • Minimiere kognitive Ablenkung. • Verhindere soziale Bedrohung.
2. **Biete Wahlmöglichkeiten bei der sprachlichen und symbolischen Darstellung von Informationen.** • Biete Hilfen zur Klärung von Begriffen und Symbolen. • Biete Hilfen zum Erkennen von Syntax und Textaufbau. • Biete Hilfen beim Lesen von geschriebenen Texten oder von mathematischen Formeln und Symbolen. • Biete Möglichkeiten zur Nutzung von Kenntnissen in anderen Sprachen. • Biete Möglichkeiten der nicht-sprachlichen Illustration von Schlüsselbegriffen.	5. **Biete Möglichkeiten im Bereich der Beherrschung instrumenteller und darstellender Fertigkeiten.** • Lasse verschiedene Arten der Kommunikation zu (geschriebenen oder gesprochenen Text, Zeichnungen, Filme, …). • Ermögliche die Nutzung von Hilfen beim Erstellen einer Antwort wie konkrete Materialien und Taschenrechner in Mathematik oder Wörterbücher, Textverarbeitungsprogramme, Spracherkennungs-software bei der Textproduktion. • Biete Hilfen bei instrumentellen Fertigkeiten an, die reduziert werden können (Mentoren, Tutoren, Software).	8. **Gib Gelegenheiten für unterstützte konzentrierte Anstrengung und ausdauerndes Lernen.** • Erhöhe die Sichtbarkeit und Bedeutsamkeit der Lehr- und Lernziele. • Variiere das Anforderungsniveau der Aufgaben und die verfügbaren Hilfen und optimiere auf diese Weise das individuelle Anforderungsniveau. • Fördere die Kommunikation und die Zusammenarbeit unter den Lernenden. • Biete formative Lernrückmeldungen mit Bezug auf die Lernzielerreichung an.
3. **Biete Wahlmöglichkeiten beim Verstehen von Informationen.** • Biete Möglichkeiten der Aktivierung oder Erarbeitung von Hintergrundinformationen an. • Biete Hilfen zum Hervorheben wichtiger Informationen, leitender Ideen oder Beziehungen an. • Biete Hilfen an, welche systematische Informationsverarbeitung anleiten. • Biete Hilfen an, die das Behalten und den Transfer des Gelernten unterstützen.	6. **Biete Wahlmöglichkeiten zur Unterstützung der exekutiven Funktionen.** • Initiiere und unterstütze geeignete Lernzielsetzungen. • Unterstütze geplantes und strategisches Arbeiten. • Erleichtere den geordneten Umgang mit Informationen und Ressourcen. • Biete Möglichkeiten zur Selbstevaluation und fördere Kompetenzen durch Hilfe und formatives Feedback.	9. **Biete Möglichkeiten und Hilfen für selbstreguliertes Lernen.** • Entwickle und fördere motivationsförderliche Ergebniserwartungen und Kontrollüberzeugungen. • Ermögliche individuelle Bewältigungsfähigkeiten und -strategien. • Biete Möglichkeiten zur eigenständigen Lernerfolgsmessung und zur reflexiven Beurteilung des eigenen Lernprozesses.

Zweck von Experimenten			
	Beschreibung	**Kodierregel**	**Ankerbeispiel**
Entwicklung von chemischen Wissen	Chemiker sind Forscher und entwickeln deshalb chemisches Wissen. Damit sind auch Entdeckungen und Technologie gemeint. Hierzu benutzen sie Experimente, um damit neue(s) Sachen/Ideen/Wissen zu entwickeln. Dieses Wissen kann dann weiter benutzt werden.	Wird kodiert, wenn beschrieben wird: – wobei neues Wissen entsteht ***oder/und*** – wie Wissen weiterverwendet wird ***oder/und*** – wie ein Erklärung mit Hilfe von Wissen gegeben wird.	„Wenn sie was machen, wo sie noch nicht, wissen was geschieht und das dann halt durchführen." „Ich weiß nicht so richtig, aber sie Experimentieren ja und (ähm) machen neue Entdeckungen und die wollen ja, zum Beispiel, was Neues haben." „Um, das Wissen von den weiterzuführen." „Weil sie ja wissen wollen- Sie wollen ja beobachten, wie die Sachen reagieren. Sie wollen halt irgendwie testen, wie es halt reagiert."
Experimente sollen funktionieren	Chemiker testen das Experiment auf seine Funktion. Dabei gehen sie unsystematisch vor. Sie probieren aus und entwickeln dadurch das Experiment weiter. Ein gutes Experiment wird als solches bewertet, wenn es funktioniert.	Wird kodiert, wenn beschrieben wird: – dass Experimente funktionieren sollen ***oder/und*** – wie vorgegangen wird, damit Experimente funktionieren.	„Sie gucken, ob die Experimente (ähm) funktionieren und was das Ergebnisse ist davon." „Also ich hätte gesagt, dass sie verschiedenen Versuche durchführen und wenn das nicht geklappt hat versuchen sie es halt nochmal. Und (ähm), wenn das dann trotzdem nicht klappt, versuchen sie das halt dann anders, bis dann klappt." „Sie experimentieren und we-/ und also bewerten und probieren es dann halt aus, ob es funktioniert oder nicht." „Also, ob es klappt. Und die- also, dass es dann halt funktioniert."

Zweck von Experimenten			
	Beschreibung	**Kodierregel**	**Ankerbeispiel**
Experimente helfen bei der technischen Entwicklung	Chemiker handeln wie ein Ingenieur und entwickeln zweckorientiert Dinge/Sachen/Technologie, die unter Umständen auf ein allgemeines Gut zielen (z.B.: „Wohl der Menschheit“)	Wird kodiert, wenn beschrieben wird: – wenn das Testen/Ausprobieren/ Experimentieren auf die Entwicklung von Sachen/Mitteln/Dingen bezogen wird ***und/oder*** – die Entwicklung von Sachen/Dingen/Mitteln auf ein allgemeines Gut (Sicherheit, Gesundheit, Wohl der Menschheit) bezogen wird.	„Sie machen halt ihre Ideen damit und halt selber herauszufinden, wie man so etwas machen. Also Allergiemittel, wie ich davor gesagt habe.“ „Sie probieren, sie wollen etwas ausprobieren, was man dann vielleicht später für Medikamente oder so etwas benutzen kann.“ „Chemiker mit Experimenten machen, glaube ich so halt, was für Menschen und Stoffe, die man für Technik oder Medikamente braucht.“ „Sie testen Sachen. (...) Also zum Beispiel für spezielle Kosmetiksachen zum Beispiel, damit keine komische chemische Hautreaktion kommt.“
Persönlicher Erfolg durch Experimente	Chemiker machen Experimente nicht nur aus intrinsischen Gründen, sondern auch aus extrinsischen. Sie wollen berühmt werden, Erfolg haben und Geld mit ihren Ergebnissen verdienen.	Wird kodiert, wenn beschrieben wird: – wenn das Vorhaben oder/und der Ausgang um ein Experiment auf den persönlichen Erfolg (Geld, Ruhm, Bekanntheit) bezogen wird.	„(...) wenn sie was Eigenes entworfen haben, können sie es durch eine Auktion vermarkten.“ „Damit sie Erfolg haben und berühmt werden.“ „Also, damit sie bekannt werden“
Experimente ermöglichen das Ausprobieren	Chemiker probieren mit Experimenten Dinge aus. So können Flüssigkeiten gemischt werden, um zu gucken was passiert. Experimente werden dabei beobachtet, aber nicht gezielt ausgewertet. Eigenschaften der Beobachtung werden beschrieben.	Wird kodiert, wenn beschrieben wird: – welche Mittel/Sachen/Dinge Chemiker im Experiment ungezielt benutzen ***und/oder*** – dass Experimente ein Ausprobieren darstellen ***und/oder*** – dass Experimente durch Beobachtungen abgeschlossen werden.	„Sie probieren jetzt die Experimente aus und gucken (...)“ „Sie mhm (überlegend) mischen, glaube ich, Sachen, ja.“ „Also Chemiker machen Experimente, um zu beobachten, was passiert.“ „Sachen, wie Gase zum Beispiel zusammennehmen und damit Experimentieren.“ „Ja, sie probieren halt so Sachen aus und Experimentieren halt so.“

Inquiry Cycle			
	Beschreibung	**Kodierregel**	**Ankerbeispiel**
Chemiker wählen die Experimentiermethoden aus	Chemiker wählen die Experimentiermethoden gezielt aus. Experimentiermethoden beschreiben die Art, wie Daten aus Experimenten gewonnen werden. Die Wahl der der Experimentiermethoden ermöglicht es, bestimmte Daten über die Beobachtung hinaus zu gewinnen.	Wird kodiert, wenn beschrieben wird: – wie Geräte/Dinge/Sachen/Methoden zweckbestimmt eingesetzt werden ***und/oder*** – wie eine Überlegung zum Experiment die Durchführung bestimmt.	„Zum Beispiel ein Virus oder so. Sonst wird das ja ganz schlecht für die Menschheit. Deswegen testen sie das an Tieren oder Affen." „Ja, dann machen sie die Experimentiermethode, also wie sie das machen und dann führen sie das Experiment durch." „(...) überlegen sie halt, was sie dafür brauchen, dann Durchführung." „(...) dann holen sie sich vielleicht Geräte herbei, woran sie es testen können, zum Beispiel Volumen und Gewicht oder so."
Chemiker planen ihre Untersuchung	Chemiker planen Experimente vorhinein. Ausgehend von der Frage oder der Idee/Vermutung/Hypothese wird die Anordnung des Experiments sowie der Einsatz Experimentiermethoden und der Einsatz von Chemikalien bestimmt.	Wird kodiert, wenn beschrieben wird: – wie Chemiker durch Planung/Überlegung zu einem Experiment kommen und/oder – wie Chemiker Fragen ***und/oder*** Idee/Vermutungen/Hypothesen nutzen, um ein Experiment zu planen ***und/oder*** – wie ein prototypischer Ablauf (Frage-Idee-Planung-Durchführung-Auswertung) eines Experiments aussieht.	„Erstmal müssen sie- (ähm) alle nachdenken, was sie überhaupt durchführen wollen." „Also, sie planen erst, wie sie Experimentieren und dann Experimentieren sie." „Sie stellen eine- sich selbst eine Frage, dann dann (ähm) testen, dann holen sie sich vielleicht Geräte herbei, woran sie es testen können, zum Beispiel Volumen und Gewicht oder so. Dann machen sie eine Feststellung: Zum Beispiel Luft ist schwerer als Gas, zum Beispiel. Dann haben sie ein Ergebnis."

Inquiry Cycle			
	Beschreibung	**Kodierregel**	**Ankerbeispiel**
Chemiker testen Ideen	Chemiker testen/überprüfen im Experiment ihre Ideen/Vermutungen/Hypothesen, die aus ihren Fragen abgelietet sind. Das Experiment muss entsprechend angeordnet sein, um eine Aussage über eine Hypothese zu ermöglichen.	Wird kodiert, wenn beschrieben wird: – dass Ideen vorläufig sind und einen Testbedarf aufweisen ***und/oder*** – dass Ideen die Durchführung leiten ***und/oder*** – dass Ideen aus Fragen stammen und/oder diese belegen	„Sie machen zuerst eine Vermutung oder schreiben auf ihre Vermutung, dann testen sie halt den Versuch und dann schreiben sie die (ähm)- was passiert ist auf.“ „Sie wollen (ähm) ihre Ideen testen. Also sie wollen gucken, ob ihre Ideen sich bewahrheiten.“ „Sie testen ihre Ideen, die sie im Kopf haben, damit sie ihre Frage belegen können.“
Chemiker werten Experimente aus	Chemiker beziehen die Daten aus dem Experiment auf die Frage und auf die Idee/Vermutung/Hypothese. Die Daten ermöglichen die Verifizierung/Falsifizierung der Idee/Hypothese/Vermutung.	Wird kodiert, wenn beschrieben wird: – wie die Auswertung auf die Frage bezogen wird ***und/oder*** – dass die Auswertung den Schluss des Experiments bildet ***und/oder*** – dass die Auswertung ergebnisoffen ist ***und/oder*** – dass Auswertung und Beobachtung zusammenhängen	„(...) und dann stellt man halt sozusagen seine Theorie oder so halt in Frage und guckt, was richtig- ob es richtig oder falsch ist.“ „und dann werten sie das ganze aus.“ „Also so, wie sie dies das vorgestellt haben und (ähm), dann kommt halt am Ende irgendwas heraus. „und ja gucken, was passiert.“
Chemiker stellen Fragen	Chemiker finden durch Neugierde Frage auf Zustände in der Welt. Die Fragen sind maßgeblich für die Entwicklung und Durchführung von Experimenten und deren Auswertung.	Wird kodiert, wenn beschrieben wird: – dass Fragen den Ausgangspunkt zu einem Experiment bilden ***und/oder*** – dass die Auswertung eine Antwort auf die Frage ist ***und/oder*** – dass Frage und Zielstellung zusammenhängen (Zweck)	„Also (ähm) erstmal brauchen sie eine Frage, die sie beantworten wollen“ „(...) damit sie ihre Frage belegen können.“ „Weil sie ihre Fragen, also die wollen herausfinden, was mit das und das machen kann glaube ich.“ „Sie stellen eine- sich selbst eine Frage (...)“

Art der Äußerungen (Generell vs. NOS-spezifisch)			
	Beschreibung	**Kodierregel**	**Ankerbeispiel**
Generelle Äußerung über den Inhalt	Die generelle Äußerung nimmt einen Bezug zum Inhalt der Frage stellt aber keinen direkten Bezug zu einer NOS-Dimension her	Wird kodiert, wenn: – der Inhalt Bezug zur Frage hat und nicht abweisend ist – keine Ideen/Hypothesen/Vermutungen genannt werden ***und/oder*** – kein Test-/Vorläufigkeitscharakter der Idee/Hypothesen/Vermutungen vorliegt	„Sie schreiben die Beobachtung und die Durchführung auf und ja gucken, was passiert." „Also ich stell so Leute vor, die mit diesen Kolben- Also deren Flaschen, überall so umkippen und daraus eine chemische Mischung entsteht." „Also zum Beispiel, wenn da was (ähm)- Wenn da irgendeine (ähm) eine Bakterie drin ist, die für den Körper nicht gut ist."
Kontextbasierte NOS-Äußerung	Die inhaltsbasierte Diskussion führen Lernende, wenn sie den Lerninhalt auf wissenschaftlichen Ideen/Hypothesen/Vermutungen beziehen.	Wird kodiert, wenn: – Ideen/Hypothesen/Vermutungen genannt werden und ihr Test- oder Vorläufigkeitscharakter genannt wird ***und/oder*** – Ideen/Hypothesen/Vermutungen mit Bezug auf einen Inhalt beschrieben werden	„Ich glaube (…), dass sie erstmal so Sach- also. Also (...) erstmal gucken, was könnte man- was. Also was regt den Wachstum der Pflanze an. Und dann könnte man damit gucken, was das- (...) ja, was halt helfen könnte." „Wie hoch eine Flamme steigen kann, oder so" „Die Menge. Ob sie gleich auch gleich viel wiegt. Und dann nimmt man einfach von verschwenden Stoffe und die gleiche Menge ab und dann wiegt man sie. (..) Und dann machen sie Beobachtung, was da passiert und dann durch diese Beobachtungen können sie sich eigentlich ihre Frage beantworten."

Art der Äußerungen (Generell vs. NOS-spezifisch)			
	Beschreibung	**Kodierregel**	**Ankerbeispiel**
Kontextlose NOS-Äußerung	Die inhaltslose Diskussion führen Lernende, wenn sie wissenschaftlichen Ideen/Hypothesen/Vermutungen oder NOS-Konzepte (Rechtfertigung, Zweck bei Kremer, soziale Eingebundenheit), direkt thematisieren, ohne auf den Lerninhalt einzugehen.	Wird kodiert, wenn: – allgemein diskutiert wird, dass Ideen/Hypothesen/Vermutungen Test- oder Vorläufigkeitscharakter aufweisen **und/oder** – beschrieben wird, wie das Experiment den Ausgangspunkt zu neuem Wissen bilden **und/oder** – beschrieben wird, dass Chemiker mit Experimenten neue Entdeckungen machen **und/oder** – beschrieben wird, dass es wichtig ist konkrete Vorstellung zu einem Experiment haben **und/oder** – beschrieben wird, dass Chemiker Dinge/Sachen erklären **und/oder** – beschrieben wird, dass bestimmte Beobachtungen/Ergebnisse ihrer Experimente erwarten **und/oder** – beschrieben wird, dass Chemiker aus sich heraus Frage und Ideen entwickeln **und/oder** – beschrieben wird, dass ein Experiment ein guter Weg ist, um herauszufinden, ob etwas wahr ist.	„Vielleicht damit, die was Neues über diese Sachen vielleicht rausfinden.“ „Sie überlegen sich erstmal was und haben eine Idee und versuchen es mit Experimenten umzusetzen.“ „Sie wollen (ähm) ihre Ideen testen. Also sie wollen gucken, ob ihre Ideen sich bewahrheiten.“ „Sie testen ihre Ideen, die sie im Kopf haben, damit sie ihre Frage belegen können.“ „Sie testen ihre Ideen, ob sie funktionieren oder nicht, weil, wenn sie es nicht testen würden, würden sie nicht wissen, ob es funktioniert oder nicht.“

7.5 Interviewleitfaden der Vorstudie

<table>
<tr><td colspan="4">Anweisungen für den Interviewer:

Sei entspannt und ruhig. Deine Aufregung überträgt sich! Es geht nicht darum, Dich zu testen! Es geht ausschließlich um die Lebenswelt des Interviewpartners! Dir muss nichts peinlich sein. Du solltest aber auch nicht übertrieben dankbar sein.
Versuche die Botschaften des Interviewpartners zu verstehen. Frag bei Unklarheiten nach! Es gibt keine naiven Fragen! Stelle aber nicht Deine eigene Position dar.
Zeig ein unabhängiges Interesse. Sei offen für alle Arten von Informationen.
Die Gruppe agiert primär selbständig:
Kein Eingriff in die Verteilung der Redebeiträge.
Aufrechterhaltung des Redeflusses. Halte dich an den Leitfaden:
Stelle leichte, kurze Fragen!
Verwende keine Fachbegriffe!
Aber: Nutze Deine Sprache. Das wirkt authentisch.
Beispiele:
Kannst Du das nochmal sagen?
Das habe ich nicht verstanden. Kannst Du das nochmal anders sagen?
</td></tr>
<tr><td colspan="4">Ablauf des Interviews:</td></tr>
<tr><td colspan="2">Briefing: Schaffe eine offene, entspannte Atmosphäre.</td><td colspan="2">1. Die Interviews dienen der Erprobung von neuen Aufgaben für den Chemieunterricht.
2. Alle Informationen werden anonym verwendet.
3. Die Interviews finden im Rahmen einer Promotion statt.
4. Die Gespräche werden mit dem Diktiergerät aufgezeichnet.
5. Aufteilung in die Raumecken. Einzelinterviews beginnen.
6. Jetzt den Recorder einschalten.</td></tr>
<tr><td rowspan="3">Einzelinterviews: NOS-Verständnis</td><td>Fragen</td><td>Alternativen</td><td>Mögliche Schülerantworten</td></tr>
<tr><td>Was machen Chemiker mit Experimenten?

Warum machen Chemiker Experimente?</td><td>Wozu nutzen Chemiker Experimente?

Wie nutzen Chemiker Experimente?

Mit welchem Ziel machen Chemiker Experimente?

Zu welchem Zweck machen Chemiker Experimente?</td><td>Chemiker machen mit Experimenten Entdeckungen.

Chemiker suchen nach neuen Dingen. Dazu nutzen sie Experimente.

Chemiker wollen Neues entdecken.</td></tr>
<tr><td>Wie gehen Chemiker vor, wenn sie experimentieren?</td><td>Wie machen Chemiker ein Experiment?

Was tun Chemiker, wenn sie experimentieren?</td><td>Sie wählen Geräte und Chemikalien und gucken was passiert.

Chemiker machen mit gefährlichen Stoffe tolle Sachen.

Chemiker fangen einfach an.</td></tr>
<tr><td colspan="4">Gruppenarbeit: Bearbeitung der Aufgaben</td></tr>
<tr><td></td><td colspan="3">Wir beginnen jetzt. Bitte bearbeitet jetzt die Aufgaben. Benutzt das Tablet. Es gibt Buttons zum Drücken zum Vorlesen oder für Hilfe. Die Lautstärke könnt ihr hier regeln. Sagt alles, was ihr denkt, laut. Arbeitet in der Gruppe und helft euch gegenseitig.</td></tr>
</table>

<table>
<tr><td colspan="4">Überleitung zu den Einzelinterviews</td></tr>
<tr><td rowspan="3">Einzelinterviews: NOS-Verständnis</td><td>Fragen</td><td>Alternativen</td><td>Mögliche Schülerantworten</td></tr>
<tr><td>Was machen Chemiker mit Experimenten?

Warum machen Chemiker Experimente?</td><td>Wozu nutzen Chemiker Experimente?

Wie nutzen Chemiker Experimente?

Mit welchem Ziel machen Chemiker Experimente?

Zu welchem Zweck machen Chemiker Experimente?</td><td>Chemiker nutzen Experimente, um Ideen/Zusammenhänge /Theorien zu testen.

Chemiker testen mit Experimenten Ideen.

Keine Entwicklung: Chemiker wollen Neues entdecken.</td></tr>
<tr><td>Wie gehen Chemiker vor, wenn sie experimentieren?</td><td>Wie machen Chemiker ein Experiment?

Was tun Chemiker, wenn sie experimentieren?</td><td>Sie haben Fragen und Ideen. Ideen sind Antworten auf die Fragen. Die Antworten testen sie im Experiment.

Chemiker haben Ideen zu Fragen. Mit Experimentiermethoden testen sie diese Ideen.

Keine Entwicklung: Chemiker fangen einfach an.</td></tr>
<tr><td colspan="4">Adaptivität der Lernaufgaben</td></tr>
<tr><td></td><td colspan="3">Was hast du mitgenommen? Was war neu für dich?

Sind Fragen offen?

Wie gefielen dir die Aufgaben selbst?

War immer klar, was du machen musst?

Warst du unter- oder überfordert?

Was hat dir geholfen?

Was hat dir nicht geholfen?</td></tr>
</table>

7.6 Statistische Beschreibung beider Assessmentversionen

Tabelle 7.1: Rel. Verteilung der Antwortkategorien pro Item im UDA-Assessment.

	Messzeitpunkt 1					Messzeitpunkt 2				
NOS-Skala	1	2	3	4	5	1	2	3	4	5
Herkunft										
Item 1	13%	6%	15%	11%	55%	11%	11%	14%	12%	52%
Item 2	6%	6%	14%	21%	54%	5%	10%	16%	25%	43%
Item 3	31%	20%	16%	18%	16%	12%	16%	28%	25%	20%
Item 4	6%	15%	20%	20%	39%	9%	10%	27%	19%	34%
Item 5	4%	3%	15%	24%	53%	10%	9%	21%	22%	38%
Sicherheit										
Item 1	3%	10%	23%	23%	41%	8%	6%	23%	21%	42%
Item 2	7%	7%	31%	30%	25%	9%	13%	35%	21%	21%
Item 3	6%	9%	19%	25%	42%	6%	9%	22%	20%	43%
Item 4	7%	7%	18%	19%	49%	11%	11%	21%	18%	39%
Item 5	5%	9%	21%	26%	39%	9%	13%	22%	25%	31%
Item 6	2%	1%	13%	27%	57%	7%	8%	25%	29%	31%
Item 7	5%	9%	17%	29%	40%	9%	14%	25%	20%	32%
Entwicklung										
Item 1	15%	25%	31%	15%	13%	7%	16%	28%	25%	23%
Item 2	10%	19%	31%	24%	17%	5%	14%	24%	25%	32%
Item 3	7%	10%	25%	33%	26%	5%	11%	20%	29%	35%
Item 4	8%	16%	33%	23%	21%	6%	14%	25%	21%	34%
Item 5	8%	14%	33%	26%	19%	5%	11%	32%	21%	31%
Item 6	8%	9%	14%	20%	49%	6%	14%	21%	17%	43%
Item 7	8%	8%	23%	23%	39%	9%	8%	23%	25%	35%
Item 8	11%	11%	31%	27%	21%	5%	9%	27%	21%	38%
Rechtfertigung										
Item 1	7%	5%	23%	26%	39%	5%	7%	23%	17%	48%
Item 2	7%	16%	29%	23%	25%	6%	9%	32%	25%	29%
Item 3	13%	13%	30%	24%	20%	7%	12%	23%	21%	36%
Item 4	7%	11%	36%	23%	23%	4%	6%	32%	28%	28%
Item 5	4%	7%	16%	25%	48%	4%	9%	18%	25%	45%
Item 6	8%	10%	36%	28%	17%	7%	8%	33%	27%	24%
Item 7	4%	11%	21%	31%	33%	4%	12%	25%	25%	34%

Anmerkung:
Die negativ gepolten Skalen (Sicherheit und Herkunft) sind umgepolt dargestellt.

Tabelle 7.2: Relative Verteilung der Antwortkategorien pro Item im Originalassessment.

	Messzeitpunkt 1					Messzeitpunkt 2				
NOS-Skala	1	2	3	4	5	1	2	3	4	5
Herkunft										
Item 1	3%	16%	39%	22%	20%	4%	18%	23%	33%	22%
Item 2	3%	14%	31%	28%	25%	3%	14%	32%	30%	22%
Item 3	16%	31%	26%	19%	9%	4%	21%	42%	23%	9%
Item 4	9%	13%	33%	26%	20%	6%	17%	40%	27%	11%
Item 5	4%	16%	35%	32%	13%	4%	15%	38%	27%	16%
Sicherheit										
Item 1	3%	15%	46%	25%	12%	2%	16%	32%	30%	19%
Item 2	2%	12%	27%	36%	23%	4%	11%	31%	27%	27%
Item 3	5%	13%	29%	38%	15%	4%	17%	31%	28%	20%
Item 4	5%	12%	36%	29%	17%	7%	14%	32%	28%	18%
Item 5	1%	6%	31%	32%	30%	3%	12%	28%	30%	27%
Item 6	3%	22%	37%	29%	10%	4%	19%	43%	20%	14%
Item 7	9%	14%	40%	28%	9%	9%	21%	38%	20%	12%
Entwicklung										
Item 1	2%	16%	34%	39%	9%	4%	6%	36%	36%	17%
Item 2	1%	8%	27%	41%	23%	4%	6%	29%	34%	27%
Item 3	4%	5%	26%	37%	27%	3%	6%	24%	39%	28%
Item 4	3%	6%	35%	35%	21%	4%	4%	25%	43%	24%
Item 5	1%	7%	38%	43%	11%	4%	6%	34%	42%	14%
Item 6	1%	4%	14%	40%	41%	4%	4%	25%	31%	36%
Item 7	1%	1%	34%	41%	23%	4%	4%	25%	44%	23%
Item 8	1%	5%	38%	45%	11%	5%	2%	35%	40%	18%
Rechtfertigung										
Item 1	1%	8%	18%	31%	42%	6%	7%	21%	37%	29%
Item 2	3%	8%	32%	46%	11%	4%	7%	23%	56%	10%
Item 3	5%	8%	32%	36%	19%	2%	7%	30%	41%	20%
Item 4	3%	7%	26%	44%	20%	2%	7%	25%	38%	28%
Item 5	NA	3%	13%	45%	40%	1%	2%	20%	34%	43%
Item 6	3%	1%	45%	41%	10%	3%	6%	36%	36%	20%
Item 7	2%	3%	29%	44%	22%	1%	4%	33%	38%	23%

Anmerkung:
Die negativ gepolten Skalen (Sicherheit und Herkunft) sind umgepolt dargestellt.

Tabelle 7.3: Statistische Beschreibungen des UDA-Assessments.

	MZP 1				MZP 2			
Item	M	SD	Schiefe	W ölbung	M	SD	Schiefe	Wölbung
Herkunft								
1	2.10	1.45	0.95	-0.58	2.16	1.43	0.80	-0.83
2	1.87	1.18	1.26	0.61	2.09	1.22	0.87	-0.32
3	3.32	1.46	-0.27	-1.35	2.74	1.27	0.24	-0.96
4	2.30	1.29	0.55	-0.95	2.42	1.30	0.48	-0.86
5	1.82	1.09	1.33	1.13	2.31	1.33	0.68	-0.72
Sicherheit								
1	2.11	1.15	0.67	-0.57	2.17	1.25	0.80	-0.37
2	2.39	1.13	0.58	-0.26	2.68	1.22	0.23	-0.77
3	2.11	1.21	0.84	-0.29	2.15	1.25	0.76	-0.51
4	2.03	1.26	0.99	-0.15	2.37	1.38	0.59	-0.93
5	2.15	1.18	0.77	-0.34	2.44	1.29	0.51	-0.84
6	1.65	0.91	1.58	2.60	2.32	1.20	0.65	-0.42
7	2.09	1.17	0.90	-0.11	2.49	1.31	0.39	-1.00
Entwicklung								
1	2.87	1.24	0.20	-0.88	3.40	1.22	-0.28	-0.88
2	3.18	1.21	-0.12	-0.88	3.66	1.20	-0.49	-0.80
3	3.60	1.17	-0.58	-0.47	3.77	1.19	-0.68	-0.51
4	3.34	1.20	-0.20	-0.82	3.63	1.24	-0.43	-0.92
5	3.33	1.18	-0.27	-0.69	3.62	1.18	-0.38	-0.77
6	3.93	1.32	-0.97	-0.32	3.77	1.29	-0.60	-0.90
7	3.78	1.25	-0.74	-0.46	3.70	1.28	-0.71	-0.53
8	3.38	1.23	-0.40	-0.70	3.78	1.19	-0.61	-0.59
Rechtfertigung								
1	3.86	1.19	-0.86	-0.06	3.96	1.20	-0.85	-0.31
2	3.43	1.23	-0.28	-0.91	3.61	1.16	-0.45	-0.57
3	3.27	1.28	-0.29	-0.89	3.67	1.27	-0.55	-0.83
4	3.45	1.16	-0.29	-0.62	3.70	1.09	-0.52	-0.28
5	4.06	1.12	-1.04	0.18	3.96	1.17	-0.91	-0.18
6	3.36	1.13	-0.35	-0.44	3.54	1.15	-0.46	-0.44
7	3.79	1.14	-0.68	-0.41	3.73	1.17	-0.53	-0.69

Anmerkung:
M: Mittelwert
SD: Standardabweichung

Tabelle 7.4: Statistische Beschreibungen des Originalsassessments.

	MZP 1				MZP 2			
Item	M	SD	Schiefe	Wölbung	M	SD	Schiefe	Wölbung
Herkunft								
1	2.61	1.06	-0.03	-0.77	2.49	1.14	0.37	-0.82
2	2.42	1.09	0.26	-0.79	2.46	1.07	0.26	-0.68
3	3.26	1.19	-0.26	-0.87	2.88	0.99	-0.07	-0.40
4	2.64	1.19	0.30	-0.70	2.81	1.05	0.15	-0.42
5	2.65	1.04	0.23	-0.50	2.64	1.04	0.10	-0.55
Sicherheit								
1	2.72	0.95	-0.07	-0.25	2.52	1.05	0.16	-0.79
2	2.34	1.03	0.40	-0.55	2.38	1.13	0.40	-0.62
3	2.56	1.07	0.47	-0.36	2.57	1.12	0.21	-0.78
4	2.60	1.08	0.29	-0.44	2.64	1.15	0.29	-0.67
5	2.15	0.95	0.32	-0.71	2.35	1.09	0.39	-0.68
6	2.79	0.98	-0.01	-0.61	2.80	1.04	-0.10	-0.48
7	2.87	1.07	0.27	-0.40	2.97	1.12	-0.04	-0.61
Entwicklung								
1	3.37	0.93	-0.25	-0.44	3.57	0.98	-0.49	0.18
2	3.79	0.92	-0.40	-0.41	3.72	1.06	-0.60	-0.09
3	3.78	1.04	-0.73	0.19	3.84	0.99	-0.72	0.21
4	3.65	0.97	-0.40	-0.07	3.79	0.97	-0.79	0.59
5	3.55	0.83	-0.32	0.18	3.56	0.95	-0.62	0.45
6	4.18	0.87	-0.97	0.67	3.91	1.05	-0.80	0.16
7	3.84	0.81	-0.15	-0.31	3.76	0.99	-0.83	0.63
8	3.60	0.78	-0.19	0.05	3.64	0.97	-0.69	0.67
Rechtfertigung								
1	4.04	1.02	-0.83	-0.15	3.78	1.12	-0.82	0.07
2	3.54	0.92	-0.63	0.38	3.60	0.92	-1.01	0.99
3	3.56	1.04	-0.51	-0.12	3.68	0.94	-0.47	-0.02
4	3.73	0.95	-0.66	0.28	3.83	0.98	-0.62	-0.03
5	4.21	0.77	-0.74	0.07	4.17	0.87	-0.80	0.13
6	3.54	0.81	-0.42	1.13	3.64	0.97	-0.41	-0.01
7	3.80	0.89	-0.60	0.50	3.76	0.90	-0.36	-0.11

Anmerkung:
M: Mittelwert
SD: Standardabweichung

Tabelle 7.5: Inter-Item-Korrelationen für beide Messzeitpunkte für die Rechtfertigungsskala.

Item-Nr.:	1	2	3	4	5	6	7
Messzeitpunkt 1							
1	NA	0.24	0.03	0.23	0.19	0.15	0.22
2	0.47	NA	0.09	0.23	0.1	0.18	0.18
3	0.42	0.37	NA	0.26	0.15	0.17	0.27
4	0.41	0.35	0.24	NA	0.13	0.12	0.27
5	0.46	0.43	0.49	0.49	NA	0.07	0.17
6	0.49	0.29	0.33	0.37	0.24	NA	0.37
7	0.44	0.44	0.45	0.56	0.48	0.49	NA
Messzeitpunkt 2							
1	NA	0.3	0.18	0.26	0.33	0.19	0.57
2	0.51	NA	0.33	0.3	0.23	0.32	0.42
3	0.4	0.28	NA	0.17	0.24	0.26	0.33
4	0.3	0.37	0.4	NA	0.29	0.28	0.51
5	0.41	0.5	0.45	0.44	NA	0.34	0.37
6	0.43	0.4	0.44	0.4	0.37	NA	0.53
7	0.49	0.56	0.45	0.41	0.58	0.44	NA

Anmerkung:
Über der Diagonalen findet sich das Original- und darunter das UDA-Assessment. Alle Korrelationen sind signifikant ($p > .05$).

Tabelle 7.6: Inter-Item-Korrelationen für beide Messzeitpunkte für die Entwicklungsskala.

Item-Nr.:	1	2	3	4	5	6	7	8
Messzeitpunkt 1								
1	NA	0.33	0.31	0.44	0.43	0.33	0.31	0.39
2	0.45	NA	0.27	0.25	0.31	0.41	0.31	0.35
3	0.52	0.48	NA	0.33	0.36	0.44	0.32	0.34
4	0.48	0.23	0.41	NA	0.35	0.37	0.53	0.44
5	0.37	0.65	0.49	0.47	NA	0.37	0.28	0.3
6	0.3	0.38	0.41	0.57	0.56	NA	0.43	0.3
7	0.38	0.4	0.45	0.47	0.36	0.47	NA	0.56
8	0.41	0.44	0.47	0.48	0.43	0.31	0.47	NA
Messzeitpunkt 2								
1	NA	0.43	0.58	0.56	0.63	0.43	0.59	0.62
2	0.69	NA	0.5	0.42	0.62	0.62	0.61	0.38
3	0.49	0.64	NA	0.53	0.64	0.56	0.6	0.57
4	0.55	0.48	0.44	NA	0.54	0.6	0.53	0.52
5	0.49	0.57	0.61	0.59	NA	0.42	0.47	0.56
6	0.51	0.5	0.56	0.53	0.45	NA	0.69	0.58
7	0.45	0.6	0.39	0.64	0.6	0.52	NA	0.69
8	0.6	0.62	0.51	0.67	0.48	0.55	0.54	NA

Anmerkung:
Über der Diagonalen findet sich das Original- und darunter das UDA-Assessment.
Alle Korrelationen sind signifikant ($p > 0.05$).

Tabelle 7.7: Inter-Item-Korrelationen für beide Messzeitpunkte für die Sicherheitsskala.

Item-Nr.:	1	2	3	4	5	6	7
Messzeitpunkt 1							
1	NA	0.25	0.39	0.25	0.29	0.24	0.4
2	0.23	NA	0.31	0.27	0.32	0.42	0.35
3	0.31	0.43	NA	0.16	0.25	0.32	0.41
4	0.2	0.38	0.41	NA	0.36	0.39	0.38
5	0.32	0.24	0.15	0.22	NA	0.48	0.37
6	0.26	0.54	0.32	0.42	0.35	NA	0.48
7	0.39	0.53	0.43	0.28	0.42	0.36	NA
Messzeitpunkt 2							
1	NA	0.5	0.54	0.51	0.39	0.35	0.5
2	0.46	NA	0.48	0.44	0.43	0.46	0.41
3	0.58	0.37	NA	0.42	0.48	0.48	0.44
4	0.4	0.48	0.47	NA	0.44	0.33	0.3
5	0.43	0.49	0.34	0.45	NA	0.39	0.34
6	0.46	0.49	0.54	0.5	0.46	NA	0.5
7	0.55	0.57	0.57	0.57	0.52	0.44	NA

Anmerkung:
Über der Diagonalen findet sich das Original- und darunter das UDA-Assessment.
Alle Korrelationen sind signifikant ($p > 0.05$).

Tabelle 7.8: Inter-Item-Korrelationen für beide Messzeitpunkte für die Herkunftsskala.

Item-Nr.:	1	2	3	4	5
Messzeitpunkt 1					
1	NA	0.29	0.3	0.19	0.39
2	0.21	NA	0.27	0.18	0.35
3	0.21	0.24	NA	0.32	0.24
4	0.21	0.12	0.15	NA	0.54
5	0.24	0.35	0.34	0.33	NA
Messzeitpunkt 2					
1	NA	0.41	0.35	0.28	0.38
2	0.6	NA	0.36	0.37	0.57
3	0.45	0.44	NA	0.47	0.43
4	0.32	0.39	0.54	NA	0.43
5	0.54	0.6	0.45	0.56	NA

Anmerkung:
Über der Diagonalen befindet sich das Original- und darunter das UDA-Assessment.
Alle Korrelationen sind signifikant ($p > .05$).

Tabelle 7.9: Verteilung absoluter und relativer fehlender Werte für das UDA-Assessment zum Messzeitpunkt 1.

	Item								
NOS-Skala	1	2	3	4	5	6	7	8	Mittelwert
Absolute Verteilung									
Herkunft	0	18	13	15	15	NA	NA	NA	12.2
Sicherheit	18	14	10	13	17	15	15	NA	14.57
Entwicklung	15	11	12	12	9	9	17	16	12.62
Rechtfertigung	42	9	9	9	11	10	12	NA	14.57
Relative Verteilung									
Herkunft	0	10.29	7.43	8.57	8.57	NA	NA	NA	6.97
Sicherheit	10.29	8	5.71	7.43	9.71	8.57	8.57	NA	8.33
Entwicklung	8.57	6.29	6.86	6.86	5.14	5.14	9.71	9.14	7.21
Rechtfertigung	24	5.14	5.14	5.14	6.29	5.71	6.86	NA	8.33

Anmerkung:
Nicht vorhandene Items sind mit NA gekennzeichnet

Tabelle 7.10: Verteilung absoluter und relativer fehlender Werte für das Originalassessment zum Messzeitpunkt 1.

	Item								
NOS-Skala	1	2	3	4	5	6	7	8	Mittelwert
Absolute Verteilung									
Herkunft	0	13	14	24	25	NA	NA	NA	15.2
Sicherheit	24	14	14	18	18	19	18	NA	17.86
Entwicklung	24	15	15	18	20	19	18	19	18.5
Rechtfertigung	19	14	16	19	18	19	20	NA	17.86
Relative Verteilung									
Herkunft	0	7.88	8.48	14.55	15.15	NA	NA	NA	9.21
Sicherheit	14.55	8.48	8.48	10.91	10.91	11.52	10.91	NA	10.82
Entwicklung	14.55	9.09	9.09	10.91	12.12	11.52	10.91	11.52	11.21
Rechtfertigung	11.52	8.48	9.7	11.52	10.91	11.52	12.12	NA	10.82

Anmerkung:
Nicht vorhandene Items sind mit NA gekennzeichnet

Tabelle 7.11: Verteilung absoluter und relativer fehlender Werte für das UDA-Assessment zum Messzeitpunkt 2.

	Item								
NOS-Skala	1	2	3	4	5	6	7	8	Mittelwert
Absolute Verteilung									
Herkunft	25	23	22	19	22	NA	NA	NA	22.2
Sicherheit	28	17	15	18	23	18	21	NA	20
Entwicklung	20	14	14	15	14	14	20	20	16.38
Rechtfertigung	38	13	14	15	17	16	18	NA	18.71
Relative Verteilung									
Herkunft	14.29	13.14	12.57	10.86	12.57	NA	NA	NA	12.69
Sicherheit	16	9.71	8.57	10.29	13.14	10.29	12	NA	11.43
Entwicklung	11.43	8	8	8.57	8	8	11.43	11.43	9.36
Rechtfertigung	21.71	7.43	8	8.57	9.71	9.14	10.29	NA	10.69

Anmerkung:
Nicht vorhandene Items sind mit NA gekennzeichnet

Tabelle 7.12: Verteilung absoluter und relativer fehlender Werte für das Originalassessment zum Messzeitpunkt 2.

	Item								
NOS-Skala	1	2	3	4	5	6	7	8	Mittelwert
Absolute Verteilung									
Herkunft	22	24	26	27	26	NA	NA	NA	25
Sicherheit	27	24	26	27	27	27	26	NA	26.29
Entwicklung	27	22	22	24	26	24	25	25	24.38
Rechtfertigung	27	22	25	27	27	27	27	NA	26
Relative Verteilung									
Herkunft	13.33	14.55	15.76	16.36	15.76	NA	NA	NA	15.15
Sicherheit	16.36	14.55	15.76	16.36	16.36	16.36	15.76	NA	15.93
Entwicklung	16.36	13.33	13.33	14.55	15.76	14.55	15.15	15.15	14.77
Rechtfertigung	16.36	13.33	15.15	16.36	16.36	16.36	16.36	NA	15.75

Anmerkung:
Nicht vorhandene Items sind mit NA gekennzeichnet

Tabelle 7.13: Standardisierte Faktorladungen der Assessmentversionen zu beiden Messzeitpunkten.

	UDA-Assessment			Originalassessment		
	MZP 1	MZP 2	Mittelwert	MZP 1	MZP 2	Mittelwert
Herkunftsskala						
Item 1	0.44	0.63	0.53	0.42	0.46	0.44
Item 2	0.27	0.77	0.52	0.5	0.66	0.58
Item 3	0.54	0.65	0.59	0.37	0.71	0.54
Item 4	0.6	0.79	0.7	0.78	0.78	0.78
Item 5	0.52	0.69	0.61	0.69	0.63	0.66
Sicherheitsskala						
Item 1	0.56	0.66	0.61	0.43	0.75	0.59
Item 2	0.42	0.63	0.52	0.44	0.65	0.55
Item 3	0.65	0.75	0.7	0.66	0.7	0.68
Item 4	0.82	0.73	0.77	0.55	0.7	0.63
Item 5	0.53	0.7	0.62	0.53	0.61	0.57
Item 6	0.64	0.71	0.67	0.63	0.63	0.63
Item 7	0.62	0.73	0.67	0.66	0.62	0.64
Entwicklungsskala						
Item 1	0.62	0.72	0.67	0.56	0.66	0.61
Item 2	0.66	0.73	0.69	0.64	0.71	0.67
Item 3	0.72	0.6	0.66	0.51	0.73	0.62
Item 4	0.66	0.8	0.73	0.65	0.76	0.7
Item 5	0.73	0.78	0.75	0.67	0.65	0.66
Item 6	0.56	0.69	0.63	0.54	0.68	0.61
Item 7	0.6	0.61	0.6	0.72	0.83	0.78
Item 8	0.6	0.7	0.65	0.75	0.79	0.77
Rechtfertigungsskala						
Item 1	0.53	0.55	0.54	0.31	0.39	0.35
Item 2	0.72	0.7	0.71	0.25	0.38	0.31
Item 3	0.52	0.6	0.56	0.21	0.46	0.33
Item 4	0.63	0.69	0.66	0.44	0.71	0.57
Item 5	0.69	0.68	0.68	0.44	0.75	0.59
Item 6	0.66	0.68	0.67	0.5	0.73	0.62
Item 7	0.69	0.68	0.68	0.58	0.57	0.58

Anmerkung:
MZP: Messzeitpunkt

7.7 Statistische Beschreibungen der Skalen zum sozioökonomischen Status, der kognitiven Aktivierung und der E-Booknutzung

Bei der Skala zum sozioökonomischen Status sind die Fragen entweder mit "Ja" oder "Nein" zu beantworten oder aber mit "Keins", "Eins" oder "Zwei oder mehr". Entsprechend werden die Punkte für die Summenscores vergeben (Tab. 7.14).

Tabelle 7.14: Statistische Beschreibungen der Skala zum sozioökonomischen Status (Torsheim et al., 2016)

Item	M	SD	Median
Wie viele Autos haben deine Eltern? (max. 2 Punkte)	1.43	0.64	2
Wie viele Computer habt ihr zu Hause? (max. 2 Punkte)	1.68	0.54	2
Wie viele Badezimmer habt ihr zu Hause? (max. 2 Punkte)	1.63	0.51	2
Habt ihr einen Geschirrspüler? (max. 1 Punkte)	0.07	0.25	0
Hast du ein eigenes Zimmer? (max. 1 Punkte)	0.14	0.34	0
Summenscore	4.56	1.63	5

Anmerkung:
M: Mittelwert
SD: Standardabweichung

Für die Skalen zur kognitiven Aktivierung und zur E-Booknutzung musste den Aussagen der Skala zugestimmt werden muss (1 = "Stimme gar nicht zu."; 5 = "Stimme völlig zu.") (Tab. 7.15 und 7.16).

Tabelle 7.15: Statistische Beschreibungen der Skala zur kognitiven Aktivierung (Fauth, 2014)

Item	M	SD	Median
Über die Aufgaben von heute musste ich gut nachdenken.	3.78	1.09	4.0
Über die Fragen von heute musste ich gut nachdenken.	3.54	1.14	4.0
Die Aufgaben von heute waren zu Beginn schwierig.	3.11	1.31	3.0
Die Aufgaben von heute habe ich gerne gemacht.	4.06	1.08	4.0
Mittelwert	3.62	0.91	3.5

Anmerkung:
M: Mittelwert
SD: Standardabweichung

Tabelle 7.16: Statistische Beschreibungen der Skala zur Wahrnehmung des E-Books (Sprague und Dahl, 2009)

Item	M	SD	Median
Ich mag es, das E-Book im Unterricht zu haben.	3.81	1.21	4.00
Durch das E-Book finde ich Natur interessanter.	3.59	1.20	4.00
Ich kann mit dem E-Book leichter lernen.	3.51	1.29	4.00
Durch das E-Book habe ich mehr Freude im Naturunterricht.	4.08	1.07	4.00
Mittelwert	3.75	1.00	3.75

Anmerkung:
M: Mittelwert
SD: Standardabweichung

7.8 Skript zu einer longitudinalen Messinvarianzprüfung im Multi-Group-Design

Im Folgenden wird die longitudinale Messinvanrinzprüfung der Vollskala der Herkunftsdimension im Multigroupdesign beschrieben (Little, 2013; Newsom, 2015). In der Listenfunktion (`c()`)) von R werden die Beschränkungen aufgeführt. Diese werden nach und nach gleichgesetzt. Hierzu wird das `lavaan`-Paket verwendet (Rosseel, 2012).

Statt einer Messinvarianzprüfung über die Formulierung von Modellen, kann auch die Funktion `measurementInvariance()` aus dem Paket `semTools` verwendet werden (Jorgensen et al., 2018). Hier sind die Untersuchungsmöglichkeiten bei Messvarianz und die Modellspezifikation rudimentär.

Das Panelmodell ergibt sich durch die Entfernung des Kovarianzoperators (~~) zwischen den latenten Konstrukten durch einen Regressionsoperator (~).

Zum Schluss zeigt das Beispiel die Umformung der Ergebnisse aus den Analysen zur internen Konsistenz (Funktion: `reliability()` aus `semTools` und `psych`) in eine automatisierte Tabelle mit `knitr` und `kableExtra` (Jorgensen et al., 2018; Revelle, 2018; Xie, 2018; Zhu, 2018).

```
# Paketinstallation - nur einmalig notwendig
install.packages("lavaan")
install.packages("knitr")
install.packages("semTools")
install.packages("kableExtra")
install.packages("psych")

# Pakete laden
library(lavaan)
library(knitr)
library(semTools)
library(kableExtra)
library(psych)

# Konfigurale Messinvarianzprüfung:
# Keine Beschränkungen zur Testform oder dem Messzeitpunkt.
# Die Reliabilitätswerte werden je am Ende bestimmt.

config <-  '
# Lantente Variablen stehen vor dem Operator (=~) manifeste Variablen
# danach.
Source1 =~ c(l1a,l1b)*Herkunft_MZP1_1 + c(l2a,l2b)*Herkunft_MZP1_2 +
c(l3a,l3b)*Herkunft_MZP1_3 + c(l4a,l4b)*Herkunft_MZP1_4 +
c(l5a,l5b)*Herkunft_MZP1_5
Source2 =~ c(l1c,l1d)*Herkunft_MZP2_1 + c(l2c,l2d)*Herkunft_MZP2_2 +
c(l3c,l3d)*Herkunft_MZP2_3 + c(l4c,l4d)*Herkunft_MZP2_4 +
```

```
c(l5c,l5d)*Herkunft_MZP2_5

# Effects Coded Method nach Little, 2013. Aufteilung der
# beobachteten Varianz auf die Anzahl der manifesten Variablen.
# Alternative zum Referentenansatz.

l1a == 5-l2a-l3a-l4a-l5a
l1b == 5-l2b-l3b-l4b-l5b
l1a == 5-l2c-l3c-l4c-l5c
l1b == 5-l2d-l3d-l4d-l5d

# Latente Varianzen
Source1~~Source1
Source2~~Source2

# Latente Mittelwerte
Source1~1
Source2~1

# Latente Kovarianzen
Source2~~Source1

# Manifeste Varianzen Messzeitpunkt 1
Herkunft_MZP1_1~~Herkunft_MZP1_1
Herkunft_MZP1_2~~Herkunft_MZP1_2
Herkunft_MZP1_3~~Herkunft_MZP1_3
Herkunft_MZP1_4~~Herkunft_MZP1_4
Herkunft_MZP1_5~~Herkunft_MZP1_5

# Manifeste Varianzen Messzeitpunkt 2
Herkunft_MZP2_1~~Herkunft_MZP2_1
Herkunft_MZP2_2~~Herkunft_MZP2_2
Herkunft_MZP2_3~~Herkunft_MZP2_3
Herkunft_MZP2_4~~Herkunft_MZP2_4
Herkunft_MZP2_5~~Herkunft_MZP2_5

# Manifeste Kovarianzen über beide Messzeitpunkte
Herkunft_MZP2_1~~Herkunft_MZP1_1
Herkunft_MZP2_2~~Herkunft_MZP1_2
Herkunft_MZP2_3~~Herkunft_MZP1_3
Herkunft_MZP2_4~~Herkunft_MZP1_4
Herkunft_MZP2_5~~Herkunft_MZP1_5

# Manifeste Mittelwerte Messzeitpunkt 1
Herkunft_MZP1_1~c(r1a,r1b)*1
Herkunft_MZP1_2~c(r2a,r2b)*1
Herkunft_MZP1_3~c(r3a,r3b)*1
Herkunft_MZP1_4~c(r4a,r4b)*1
```

```
Herkunft_MZP1_5~c(r5a,r5b)*1

# Manifeste Mittelwerte Messzeitpunkt 2
Herkunft_MZP2_1~c(r1c,r1d)*1
Herkunft_MZP2_2~c(r2c,r2d)*1
Herkunft_MZP2_3~c(r3c,r3d)*1
Herkunft_MZP2_4~c(r4c,r4d)*1
Herkunft_MZP2_5~c(r5c,r5d)*1

# Effects Coded Method nach Little, 2013.
# Bisher keine Tau-Äquivalenz über beide Messzeitpunkte etabliert.
r1a==0-r2a-r3a-r4a-r5a
r1b==0-r2b-r3b-r4b-r5b
r1c==0-r2c-r3c-r4c-r5c
r1d==0-r2d-r3d-r4d-r5d
'

fit.konfigMI <- lavaan(model = config,data = DATA,
                                group = "Testform",
                                estimator = "mlr",
                                missing = "FIML")
summary(fit.konfigMI, standardized=TRUE,
        fit.measures=TRUE, rsquare = TRUE)

rel.list <- list()

# Gibt die internen Konsistenzen dieses Modells aus und speichert
# diese als Listenelement.
rel.list$config <- reliability(fit.konfigMI)

# Metrische Messinvarianzprüfung: Beschränkungen der Faktorladungen
# über die Testform und den Messzeitpunkt. Im Beispiel ändern sich
# die Beschränkungsterme: z.B. "l1a","l1b","l1c","l1d" zu "l1".

metrisch <-  '

Source1 =~ c(l1,l1)*Herkunft_MZP1_1 + c(l2,l2)*Herkunft_MZP1_2 +
c(l3,l3)*Herkunft_MZP1_3 + c(l4,l4)*Herkunft_MZP1_4 +
c(l5,l5)*Herkunft_MZP1_5
Source2 =~ c(l1,l1)*Herkunft_MZP2_1 + c(l2,l2)*Herkunft_MZP2_2 +
c(l3,l3)*Herkunft_MZP2_3 + c(l4,l4)*Herkunft_MZP2_4 +
c(l5,l5)*Herkunft_MZP2_5

## Effects Coded Method nach Little, 2013.
l1 == 5-l2-l3-l4-l5

# Latente Varianzen
Source1~~Source1
```

```
Source2~~Source2

# Latente Mittelwerte
Source1~1
Source2~1

# Latente Kovarianzen
Source2~~Source1

# Manifeste Varianzen Messzeitpunkt 1
Herkunft_MZP1_1~~Herkunft_MZP1_1
Herkunft_MZP1_2~~Herkunft_MZP1_2
Herkunft_MZP1_3~~Herkunft_MZP1_3
Herkunft_MZP1_4~~Herkunft_MZP1_4
Herkunft_MZP1_5~~Herkunft_MZP1_5

# Manifeste Varianzen Messzeitpunkt 2
Herkunft_MZP2_1~~Herkunft_MZP2_1
Herkunft_MZP2_2~~Herkunft_MZP2_2
Herkunft_MZP2_3~~Herkunft_MZP2_3
Herkunft_MZP2_4~~Herkunft_MZP2_4
Herkunft_MZP2_5~~Herkunft_MZP2_5

# Manifeste Kovarianzen über beide Messzeitpunkte
Herkunft_MZP2_1~~Herkunft_MZP1_1
Herkunft_MZP2_2~~Herkunft_MZP1_2
Herkunft_MZP2_3~~Herkunft_MZP1_3
Herkunft_MZP2_4~~Herkunft_MZP1_4
Herkunft_MZP2_5~~Herkunft_MZP1_5

# Manifeste Mittelwerte Messzeitpunkt 1
Herkunft_MZP1_1~c(r1a,r1b)*1
Herkunft_MZP1_2~c(r2a,r2b)*1
Herkunft_MZP1_3~c(r3a,r3b)*1
Herkunft_MZP1_4~c(r4a,r4b)*1
Herkunft_MZP1_5~c(r5a,r5b)*1

# Manifeste Mittelwerte Messzeitpunkt 2
Herkunft_MZP2_1~c(r1c,r1d)*1
Herkunft_MZP2_2~c(r2c,r2d)*1
Herkunft_MZP2_3~c(r3c,r3d)*1
Herkunft_MZP2_4~c(r4c,r4d)*1
Herkunft_MZP2_5~c(r5c,r5d)*1

# Effects Coded Method nach Little, 2013.
# Bisher keine Tau-Äquivalenz über beide Messzeitpunkte etabliert.
r1a==0-r2a-r3a-r4a-r5a
r1b==0-r2b-r3b-r4b-r5b
```

```
r1c==0-r2c-r3c-r4c-r5c
r1d==0-r2d-r3d-r4d-r5d
'

fit.metrischMI <- lavaan(model = metrisch,data = DATA,
                                    group = "Testform",
                                    estimator = "mlr")
summary(fit.metrischMI, standardized=TRUE,
        fit.measures=TRUE, rsquare = TRUE)

# Gibt die internen Konsistenzen dieses Modells aus und speichert
# diese als Listenelement.
rel.list$metrisch <- reliability(fit.metrischMI)

# Skalare Messinvarianzprüfung: Beschränkungen der Faktorladungen
# über die Testform und die Messzeitpunkte sowie für die Mittelwerte
# der manifesten Variablen. Im Beispiel ändern sich die
# Beschränkungsterme: z.B. "r1a","r1b","r1c","r1d" zu "r1".

skalar <-  '

Source1 =~ c(l1,l1)*Herkunft_MZP1_1 + c(l2,l2)*Herkunft_MZP1_2 +
c(l3,l3)*Herkunft_MZP1_3 + c(l4,l4)*Herkunft_MZP1_4 +
c(l5,l5)*Herkunft_MZP1_5
Source2 =~ c(l1,l1)*Herkunft_MZP2_1 + c(l2,l2)*Herkunft_MZP2_2 +
c(l3,l3)*Herkunft_MZP2_3 + c(l4,l4)*Herkunft_MZP2_4 +
c(l5,l5)*Herkunft_MZP2_5

# Effects Coded Method nach Little, 2013.
l1 == 5-l2-l3-l4-l5

# Latente Varianzen
Source1~~Source1
Source2~~Source2

# Latente Mittelwerte
Source1~1
Source2~1

# Latente Kovarianzen
Source2~~Source1

# Manifeste Varianzen zum Messzeitpunkt 1
Herkunft_MZP1_1~~Herkunft_MZP1_1
Herkunft_MZP1_2~~Herkunft_MZP1_2
Herkunft_MZP1_3~~Herkunft_MZP1_3
Herkunft_MZP1_4~~Herkunft_MZP1_4
Herkunft_MZP1_5~~Herkunft_MZP1_5
```

```
# Manifeste Varianzen zum Messzeitpunkt 2
Herkunft_MZP2_1~~Herkunft_MZP2_1
Herkunft_MZP2_2~~Herkunft_MZP2_2
Herkunft_MZP2_3~~Herkunft_MZP2_3
Herkunft_MZP2_4~~Herkunft_MZP2_4
Herkunft_MZP2_5~~Herkunft_MZP2_5

# Manifeste Kovarianzen über beide Messzeitpunkte
Herkunft_MZP2_1~~Herkunft_MZP1_1
Herkunft_MZP2_2~~Herkunft_MZP1_2
Herkunft_MZP2_3~~Herkunft_MZP1_3
Herkunft_MZP2_4~~Herkunft_MZP1_4
Herkunft_MZP2_5~~Herkunft_MZP1_5

# Manifeste Mittelwerte
Herkunft_MZP1_1~c(r1,r1)*1
Herkunft_MZP1_2~c(r2,r2)*1
Herkunft_MZP1_3~c(r3,r3)*1
Herkunft_MZP1_4~c(r4,r4)*1
Herkunft_MZP1_5~c(r5,r5)*1

# Effects Coded Method nach Little, 2013.
# Tau-Äquivalenz über beide Messzeitpunkte etabliert.
Herkunft_MZP2_1~c(r1,r1)*1
Herkunft_MZP2_2~c(r2,r2)*1
Herkunft_MZP2_3~c(r3,r3)*1
Herkunft_MZP2_4~c(r4,r4)*1
Herkunft_MZP2_5~c(r5,r5)*1

r1==0-r2-r3-r4-r5
'

fit.skalarMI <- lavaan(model = skalar,data = DATA,
                                group = "Testform",
                                estimator = "mlr")
summary(fit.skalarMI, standardized=TRUE,
        fit.measures=TRUE, rsquare = TRUE)

# Gibt die internen Konsistenzen dieses Modells aus und speichert
# diese als Listenelement.
rel.list$skalar <- reliability(fit.skalarMI)

# Strikte Messinvarianzprüfung: Beschränkungen der Faktorladungen
# über die Testform und die Messzeitpunkte, für die Mittelwerte der
# manifesten Variablen und deren Varianzen zu beiden Messzeitpunkten
# aber nicht über beide Messzeitpunkte. Im Beispiel werden die
# Beschränkungsterme v1 und so weiter eingeführt.
```

```
strikt <-  '
Source1 =~ c(l1,l1)*Herkunft_MZP1_1 + c(l2,l2)*Herkunft_MZP1_2 +
c(l3,l3)*Herkunft_MZP1_3 + c(l4,l4)*Herkunft_MZP1_4 +
c(l5,l5)*Herkunft_MZP1_5
Source2 =~ c(l1,l1)*Herkunft_MZP2_1 + c(l2,l2)*Herkunft_MZP2_2 +
c(l3,l3)*Herkunft_MZP2_3 + c(l4,l4)*Herkunft_MZP2_4 +
c(l5,l5)*Herkunft_MZP2_5

## Effects Coded Method nach Little, 2013.
l1 == 5-l2-l3-l4-l5

# Latente Varianzen
Source1~~Source1
Source2~~Source2

# Latente Mittelwerte
Source1~1
Source2~1

# Latente Kovarianzen
Source2~~Source1

# Manifeste Varianzen zum Messzeitpunkt 1
Herkunft_MZP1_1~~c(v1,v1)*Herkunft_MZP1_1
Herkunft_MZP1_2~~c(v2,v2)*Herkunft_MZP1_2
Herkunft_MZP1_3~~c(v3,v3)*Herkunft_MZP1_3
Herkunft_MZP1_4~~c(v4,v4)*Herkunft_MZP1_4
Herkunft_MZP1_5~~c(v5,v5)*Herkunft_MZP1_5

# Manifeste Varianzen zum Messzeitpunkt 2
Herkunft_MZP2_1~~c(v1,v1)*Herkunft_MZP2_1
Herkunft_MZP2_2~~c(v2,v2)*Herkunft_MZP2_2
Herkunft_MZP2_3~~c(v3,v3)*Herkunft_MZP2_3
Herkunft_MZP2_4~~c(v4,v4)*Herkunft_MZP2_4
Herkunft_MZP2_5~~c(v5,v5)*Herkunft_MZP2_5

# Manifeste Kovarianzen über beide Messzeitpunkte
Herkunft_MZP2_1~~Herkunft_MZP1_1
Herkunft_MZP2_2~~Herkunft_MZP1_2
Herkunft_MZP2_3~~Herkunft_MZP1_3
Herkunft_MZP2_4~~Herkunft_MZP1_4
Herkunft_MZP2_5~~Herkunft_MZP1_5

# Effects Coded Method nach Little, 2013.
# Tau-Äquivalenz über beide Messzeitpunkte etabliert.
Herkunft_MZP1_1~c(r1,r1)*1
Herkunft_MZP1_2~c(r2,r2)*1
```

```
Herkunft_MZP1_3~c(r3,r3)*1
Herkunft_MZP1_4~c(r4,r4)*1
Herkunft_MZP1_5~c(r5,r5)*1

Herkunft_MZP2_1~c(r1,r1)*1
Herkunft_MZP2_2~c(r2,r2)*1
Herkunft_MZP2_3~c(r3,r3)*1
Herkunft_MZP2_4~c(r4,r4)*1
Herkunft_MZP2_5~c(r5,r5)*1

r1==0-r2-r3-r4-r5
'

fit.striktMI <- lavaan(model = strict,data = DATA,
                                group = "Testform",
                                estimator = "mlr")
summary(fit.striktMI, standardized=TRUE,
        fit.measures=TRUE, rsquare = TRUE)

# Gibt die internen Konsistenzen dieses Modells aus und speichert
# diese als Listenelement.
rel.list$strikt <- reliability(fit.striktMI)

# Im Folgenden werden die Listenelemente aufgerufen und in einem
# Dataframe mit ribind() arrangiert.

df <- rbind(rel.list$config$UDA,
            rel.list$metrisch$UDA,
            rel.list$skalar$UDA,
            rel.list$strikt$UDA,
            rel.list$config$Original,
            rel.list$metrisch$Original,
            rel.list$skalar$Original,
            rel.list$strikt$Original)

# Das Skript erzeugt eine Tabelle mit den Ergebnissen der
# reliability()-Funktion.
# Die Zahlen in der cbind()-Funktion geben die Zeilen an,
# in denen die Omegawerte nach McDonald stehen.
# kable() erzeugt die Tabelle.

kable(cbind(c("Konfigurale MI","Metrisch MI","Skalare MI"
              ,"Strikte MI","Konfigurale MI","Metrisch MI",
              "Skalare MI","Strikte MI"),
            round(df[c(4,9,14,19,24,29,34,39),],2)),
      row.names = FALSE,format = "latex",
      col.names = c("NOS-Skala","Messzeitpunkt 1","
                    Messzeitpunkt 2",
```

```
                    "Beide Messzeitpunkte"),caption = "Interne
      Konsistenzen (McDonalds omega) bei vorgegebener
      Skalenstruktur der Herkunftsskalar.", booktabs = TRUE) %>%
  kableExtra::kable_styling("striped") %>%
  kableExtra::group_rows(group_label = "UDA-Assessment",1,4) %>%
  kableExtra::group_rows(group_label = "Originalassessment",5,8)  %>%
  footnote(general = c("MI: Messinvarianz"),
           general_title = "Anmerkung: ")
```

7.9 Skript zu einem latenten Wachstumsmodell im Multi-Group-Design

Die Etablierung von latenten Wachstumsmodellen setzt mindestens partial-skalare Messinvarianz voraus. Latente Wachstumsmodelle weisen eine second-order Struktur auf (Grimm et al., 2016; Wu et al., 2010). Das heißt, es werden weitere latente Varibalen etabliert, die durch die vorherigen modelliert sind. Im Beispiel werden die Variablen `Source1` und `Source2` genutzt, um den longitudinalen Mittelwert (`int`) und die Änderungsrate (`slope`) zu modellieren. Hier wird die Kurzskala der Herkunftsdimension verwendet.

Auch für latente Wachstumsmodelle gibt es eine spezielle Funktion (`growth()`). Diese nimmt bestimmte Modellspezifikationen selbst vor. Die Automatisierung kann jedoch stören, wenn partiale Messinvarianz vorliegt. Deshalb benutzt das Besipiel die `lavaan()`-Funktion, um das Modell völlig selbststsändig zu spezifizieren.

```
# Paketinstallation - nur einmalig notwendig
install.packages("lavaan")

# Pakete laden
library(lavaan)

Wachtum.Herk <-  '
Source1 =~ c(l1,l1)*Herkunft_MZP1_1 + c(l2,l2)*Herkunft_MZP1_2 +
c(l3,l3)*Herkunft_MZP1_3 + c(l4,l4)*Herkunft_MZP1_4 +
c(l5,l5)*Herkunft_MZP1_5
Source2 =~ c(l1,l1)*Herkunft_MZP2_1 + c(l2,l2)*Herkunft_MZP2_2 +
c(l3,l3)*Herkunft_MZP2_3 + c(l4,l4)*Herkunft_MZP2_4 +
c(l5,l5)*Herkunft_MZP2_5

## Effects Coded Method nach Little, 2013.
l2 == 4-l3-l4-l5

# Latente Varianzen werden mit psi gleichgesetzt
Source1~~psi*Source1
Source2~~psi*Source2
```

```
# Die latenten Mittelwerte werden nicht freigeschätzt, sondern mit
# 0 fixiert.
Source1~0*1
Source2~0*1

# Die Latente Kovarianz wird nicht freigeschätzt, sondern fixiert.
Source2~~0*Source1

# Manifeste Varianzen werden frei und ohne Beschränkung geschätzt.
Herkunft_MZP1_2~~Herkunft_MZP1_2
Herkunft_MZP1_3~~Herkunft_MZP1_3
Herkunft_MZP1_4~~Herkunft_MZP1_4
Herkunft_MZP1_5~~Herkunft_MZP1_5

Herkunft_MZP2_2~~Herkunft_MZP2_2
Herkunft_MZP2_3~~Herkunft_MZP2_3
Herkunft_MZP2_4~~Herkunft_MZP2_4
Herkunft_MZP2_5~~Herkunft_MZP2_5

Herkunft_MZP2_2~~Herkunft_MZP1_2
Herkunft_MZP2_3~~Herkunft_MZP1_3
Herkunft_MZP2_4~~Herkunft_MZP1_4
Herkunft_MZP2_5~~Herkunft_MZP1_5

# Effects Coded Method nach Little, 2013.
# Tau-Äquivalenz über beide Messzeitpunkte etabliert.
Herkunft_MZP1_1~c(r1,r1)*1
Herkunft_MZP1_2~c(r2,r2)*1
Herkunft_MZP1_3~c(r3,r3)*1
Herkunft_MZP1_4~c(r4,r4)*1
Herkunft_MZP1_5~c(r5,r5)*1

Herkunft_MZP2_1~c(r1,r1)*1
Herkunft_MZP2_2~c(r2,r2)*1
Herkunft_MZP2_3~c(r3,r3)*1
Herkunft_MZP2_4~c(r4,r4)*1
Herkunft_MZP2_5~c(r5,r5)*1

r2==0-r3-r4-r5

# second-order Basismodell
int =~ 1*Source1+1*Source2
slope =~ 0*Source1+1*Source2

# Latente Mittelwertde der second-order Konstrukte
int~1
slope~1
```

```
# Latente Varianzen und Kovarianzen
# der Second-order Konstrukte
slope~~slope
int~~int
slope~~int
'

fit.Wachstum.Herkunft <- lavaan(model = Wachstum.Herk,
                                data = DATA, group =
                                "Testform", estimator = "mlr",
                                missing = "FIML")
```

7.10 Skript zur DIF-Analyse

Für die DIF-Analyse mit `pairwise` ist es notwendig, dass die Daten aus den Likert-Skalen mit Null beginnend kodiert werden (Heine, 2017). Im Beispiel sind dies fünf. Außerdem zeigt das Beispiel den automatisierten Export von vier DIF-Analysen in einer kombinierten Grafik im pdf-Format sowie das Aufrufen des Exports in `R-markdown` oder `bookdown`.

```
# Paketinstallation - nur einmalig notwendig
install.packages("pairwise")
install.packages("car")
install.packages("knitr")

# Pakete laden
library(pairwise)
library(car)
library(knitr)

# Umkodierung erfolgt mit der recode()-Funktion; eine negative
# Formulierung der Items
# wird umkodiert (aus 5 wird 0 usw.).
# Die lapply() Funktion sorgt dafür, dass dies für jede
# Spalten im Datensatz passiert.
# Im Beispiel sind die negativ formulierten Items in den
# Spalten 1-12 und 28-39.

new.df <- data.frame(lapply(X = df[,c(1:12,28:39)],
                      FUN = function(x)
                      recode(x,"5=0;4=1;3=2;2=3;1=4")))

# Erzeuge vier DIF-Analysen in einer Liste mit der grm()-Funktion aus
# dem pairwise-Paket.
# nsample: XXX ist die Anzahl der Gesamtstichprobe
# m: Anzahl der Kategorien
```

```
# Es weden alle Objekte in einer Liste gespeichert.
new.df.list <- list()
new.df.list$sex <- grm(daten=new.df, m=5,
                       splitcrit = new.df$sex, nsample = XXX)
new.df.list$ses <- grm(daten=new.df, m=5,
                       splitcrit = new.df$ses, nsample = XXX)
new.df.list$kft <- grm(daten=new.df, m=5,
                       splitcrit = new.df$kft, nsample = XXX)
new.df.list$sls <- grm(daten=new.df, m=5,
                       splitcrit = new.df$sls, nsample = XXX)

# Im nächsten Schritt erzeugt das Skript vier skalierbare
# (Vektor-)Grafiken im pdf-Format.
# Als Schriftart wird Times New Roman gewählt.
# Die Funktion  par(mfrow=c(2,2)) gibt die Anordnung der Grafiken
# vor: 2 Spalten und 2 Zeilen.
# Die Anordnung erfolgt in der Reihenfolge der Grafikerzeugung.
# Die Funktion dev.off() führt den Schreibvorgang aus. Die Datei
# wird im Projektordner abgelegt. \n erzeugt eine neue Zeile im
# Title.

pdf(file='Beispiel.pdf',family = "Times New Roman")
par(mfrow=c(2,2))
plot(new.df.list$sex), xlab = "Männlich", ylab = "Weiblich",
     main = "Geschlecht", itemNames = TRUE,xymin = -2, xymax = 2)
plot(new.df.list$ses, xlab = "niedriger als Cut-off-Wert",
     ylab = "höher als Cut-off-Wert", main = "Lesefähigkeit",
     itemNames = TRUE,xymin = -2, xymax = 2)
plot(new.df.list$sls, xlab = "niedriger als Cut-off-Wert",
     ylab = "höher als Cut-off-Wert", main = "Sozioökonomischer \nStatus",
     itemNames = TRUE,xymin = -2, xymax = 2)
plot(new.df.list$kft, xlab = "niedriger als Cut-off-Wert",
     ylab = "höher als Cut-off-Wert", main = "Intelligenz",
     itemNames = TRUE,xymin = -2, xymax = 2)
dev.off()

# Mit der Funktion include_graphics() wird die Grafik in
# Bookdown/Markdown aufgerufen.
# dpi: Bestimmung der Auflösung. In dieser Dissertation wurden immer
# 300 dpi gewählt.
# path gibt den Dteipfad zur Grafik an.

include_graphics(path = 'Beispiel.pdf',dpi = 300)
```

8 | Literaturverzeichnis

Abd-El-Khalick, F. & Lederman, N. G. (2000). Improving science teachers' conceptions of nature of science: a critical review of the literature. *International Journal of Science Education*, *22* (7), 665–701. doi:10.1080/09500690050044044

Abels, S. (2015). Scaffolding inquiry-based science and chemistry education in inclusive classrooms. In N.L. Yates (Hrsg.), *New Developments in Science Education Research* (S. 77–95). Nova Science Publishers, Inc.

Abels, S. & Markic, S. (2013). Umgang mit Vielfalt – neue Perspektiven im Chemieunterricht. *Naturwissenschaften im Unterricht - Chemie*, *24* (135), 2–6.

AERA, APA & NCEO. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association & National Council on Measurement in Education.

Aikenhead, G. S. (1979). Using Qualitative Data in Formative Evaluation. *Alberta Journal of Educational Research*, *25* (2), 117–129. ERIC.

Ainscow, M. (2007). Taking an inclusive turn. *Journal of Research in Special Educational Needs*, *7* (1), 3–7. doi:10.1111/j.1471-3802.2007.00075.x

Akerson, V. L. & Hanuscin, D. L. (2007). Teaching nature of science through inquiry: Results of a 3-year professional development program. *Journal of Research in ScienceTeaching*, *44*, 653–680.

Al-Azawei, A., Serenelli, F. & Lundqvist, K. (2016). Universal Design for Learning (UDL): A Content Analysis of Peer- Reviewed Journal Papers from 2012 to 2015. *Journal of the Scholarship of Teaching and Learning*, *16* (10), 39–56. doi:10.14434/josotl.v16i3.19295

Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J. et al. (2016). *rmarkdown: Dynamic Documents for R*. Verfügbar unter: https://cran.r-project.org/package=rmarkdown

Altrichter, H., Trautmann, M., Wischer, B., Sommerauer, S. & Doppler, B. (2009). Unterrichten in heterogenen Gruppen: Das Qualitätspotenzial von Individualisierung, Differenzierung und Klassenschülerzahl. *Nationaler Bildungsbericht Österreich*, *2*, 341–360.

Anderson, D., Lai, C. F., Alonzo, J. & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment*, *16* (1), 15–34. doi:10.1080/10627197.2011.551084

Apple Inc. (2017). iBooks Author: Per Drag & Drop ist das Buch schnell erstellt. Zugriff am 15.11.2017. Verfügbar unter: https://www.apple.com/de/ibooks-author/

Autorengruppe Bildungsberichterstattung. (2016). *Bildung in Deutschland 2016*. Bielefeld: W. Bertelsmann Verlag.

Autorengruppe Bildungsberichterstattung. (2018). *Bildung in Deutschland kompakt 2018*. Bielefeld: W. Bertelsmann Verlag.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16* (3), 441. Springer.

Beauftragte der Bundesregierung für die Belange behinderter Menschen. (2014). Die UN-Behindertenrechtskonvention (UN-BKR). Bonn: Hausdruckerei BMAS.

Beddow, P. (2011). Beyond Universal Design: Accessibility Theory to Advance Testing for All Students. In *Assessing Students in the Margin: Challenges, Strategies and Techniques* (S. 581–406). Charlotte, NC: Information Age Publishing.

Bell, R. L., Mulvey, B. K. & Maeng, J. L. (2016). Outcomes of Nature of Science Instruction along a Context Continuum: Preservice Secondary Science Teachers' Conceptions and Instructional Intentions. *International Journal of Science Education*, *38* (3), 493–520. doi:10.1080/09500693.2016.1151960

Bentler, P. M. & Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociological Methods & Research*, *21* (2), 259–282. SAGE PERIODICALS PRESS.

Bernholt, S., Neumann, K. & Nentwig, P. (2012). *Making it Tangible - Specifying Learning Outcomes in Science Education*. Münster & New York: Waxmann.

Bibliographisches Institut. (2018). Wörterbuch. Zugriff am 27.7.2018. Verfügbar unter: https://www.duden.de/suchen/dudenonline/

Bridgeman, B., Trapani, C. & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41* (4), 291–310. doi:10.1111/j.1745-3984.2004.tb01167.x

Browder, D. M., Trela, K., Courtade, G. R., Jimenez, B. A., Knight, V. & Flowers, C. (2012). Teaching Mathematics and Science Standards to Students With Moderate and Severe Developmental Disabilities. *The Journal of Special Education*, *46* (1), 26–35. doi:10.1177/0022466910369942

Brown, T. (2015). *Confirmatory factor analysis for applied research* (2. Auflage). Guilford Publications. doi:10.1198/tas.2008.s98

Brownell, M. T., Smith, S. J., Crockett, J. B. & Griffin, C. C. (2012). *Inclusive instruction EvidenceBased Practices for Teaching Students with Disabilities* (S. 65–90). NewYork & London: The Guilford Press.

Burkschat, M., Cramer, E. & Kamps, U. (2012). *Beschreibende Statistik: Grundlegende Methoden der Datenanalyse* (2. Auflage). Berlin Heidelberg: Springer Spektrum.

Capp, M. J. (2017). The effectiveness of universal design for learning: a meta-analysis of literature between 2013 and 2016. *International Journal of Inclusive Education*, *21* (8), 791–807. Taylor & Francis. doi:10.1080/13603116.2017.1325074

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* (5), 514–529. doi:10.1080/0950069890110504

CAST. (2018). Universal Design for Learning (UDL) Guidelines version 2.2. Wakefield, MA: Author. Verfügbar unter: http://udlguidelines.cast.org

Chen, S. (2006). Development of an instrument to assess views on nature of science and attitudes toward teaching science. *Science Education*, *90* (5), 803–819. doi:10.1002/sce.20147

Clark, R. C. & Mayer, R. E. (2016). Introduction: Getting the most from this resource. In *e-learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning* (S. 1–6). Wiley.

Clough, M. P. (2006). Learners' Responses to the Demands of Conceptual Change: Considerations for Effective Nature of Science Instruction. *Science & Education1. Clough, M.P. (2006) Learners' Responses to the Demands of Conceptual Change: Considerations for Effective Nature of Science Instruction. Sci. Educ., 15 (5), 463–494.*, *15* (5), 463–494. doi:10.1007/s11191-005-4846-7

Conley, A. M., Pintrich, P. R., Vekiri, I. & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29* (2), 186–204. doi:10.1016/j.cedpsych.2004.01.004

Cormier, D. C., Altman, J., Shyyan, V. & Thurlow, M. L. (2010). A Summary of the Research on the Effects of Test Accommodations: 2007-2008. Technical Report 56. *National Center on Educational Outcomes, University of Minnesota*. ERIC.

Crumb, G. (1965). Understanding of science in high school physics. *Journal of Research in Science Teaching*, *3*, 246–250.

Demuth, R., Gräsel, C. & Parchmann, I. (2008). *Chemie im Kontext: von der Innovation zur nachhaltigen Verbreitung eines Unterrichtskonzepts*. Waxmann.

Deng, F., Chen, D.-T., Tsai, C.-C. & Chai, C. S. (2011). Students' views of the nature of science: A critical review of research. *Science Education*, *95* (6), 961–999. doi:10.1002/sce.20460

Deshon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46* (1), 137–149.

Deutscher, V. & Winther, E. (2018). Instructional sensitivity in vocational education. *Learning and Instruction*, *53*, 21–33. Elsevier Ltd. doi:10.1016/j.learninstruc.2017.07.004

Dorans, N. J. & Middleton, K. (2012). Addressing the Extreme Assumptions of Presumed Linkings. *Journal of Educational Measurement*, *49* (1), 1–18.

Driver, R., Leach, J. & Millar, R. (1996). *Young people's images of science*. McGraw-Hill Education (UK).

Edyburn, D. L. (2010). Would you recognize universal design for learning if you saw it? Ten propositions for new directions for the second decade of UDL. *Learning Disability Quarterly*, *33* (Winter), 33–41. doi:10.1177/073194871003300103

Elder, A. D. (2002). Characterizing fifth grade students' epistemological beliefs in science. *Personal epistemology: The psychology of beliefs about knowledge and knowing*, 347–364. Erlbaum Mahwah, NJ.

Elliott, S. N., Kettler, R. J., Beddow, P. A. & Kurz, A. (2018). *Handbook of Accessible Instruction and Testing Practices*. Cham: Springer. doi:10.1007/978-3-319-71126-3

Falk, M., Hain, J., Marohn, F., Fischer, H. & Michel, R. (2014). *Statistik in Theorie und Praxis*. Berlin, Heidelberg: Springer Spektrum. doi:10.1007/978-3-642-55253-3

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. doi:10.1016/j.learninstruc.2013.07.001

Fischer, C., Rott, D. & Veber, M. (2014). Diversität von SchülerInnen als mögliche Ressource für individuelles und wechselseitiges Lernen im Unterricht. *Lehren & Lernen*, *8* (9), 22–28.

Fuchs, L. S., Fuchs, D. & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, *37* (6), 1. Love Publishing Company.

Gesellschaft für Fachdidaktik. (2017). *Position der Gesellschaft für Fachdidaktik zum inklusiven Unterricht unter fachdidaktischer Perspektive Auf dem Weg zu inklusiven Fachdidaktiken*. (S. 1–8). Gesellschaft für Fachdidaktik.

Göransson, K. & Nilholm, C. (2014). Conceptual diversities and empirical shortcomings – a critical analysis of research on inclusive education. *European Journal of Special Needs Education*, *29* (3), 265–280. doi:10.1080/08856257.2014.933545

Greene, J. A., Cartiff, B. M. & Duke, R. F. (2018). A Meta-Analytic Review of the Relationship Between Epistemic Cognition and Academic Achievement. *Journal of Educational Psychology*. doi:10.1037/edu0000263

Grimm, K. J., Ram, N. & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Grosche, M. (2015). Was ist Inklusion? In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H. Pant Anand & M. Prenzel (Hrsg.), *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (S. 17–40). Wiesbaden: Springer Fachmedien. doi:10.1017/CBO9781107415324.004

Haladyna, T. M., State, A. & Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, *23* (1), 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x

Hall, T. E., Meyer, A. & Rose, D. (Hrsg.). (2012). *Universal Design for Learningn in the Classroom: Practical Applications*. New York & London: The Guilford Press.

Harrison, G. M., Duncan Seraphin, K., Philippoff, J., Vallin, L. M. & Brandon, P. R. (2015). Comparing Models of Nature of Science Dimensionality Based on the Next Generation Science Standards. *International Journal of Science Education*, *37* (8), 1321–1342. doi:10.1080/09500693.2015.1035357

Hattie, J. (2014). *Lernen sichtbar machen* (1. Auflage). Hohengehren: Schneider Verlag.

Heine, J.-H. (2017). pairwise: Rasch Model Parameters by Pairwise Algorithm. Verfügbar unter: http://cran.r-project.org/package=pairwise[05.01.2018]

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

Helmke, A. (2009). Unterrichtsqualität und Lehrerprofessionalität. *Diagnose, Evaluation und Verbesserung des Unterrichts*, *2*.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität : Diagnose, Evaluation und Verbesserung des Unterrichts ; Franz Emanuel Weinert gewidmet* (S. 414). Seelze-Velber: Klett.

Helmke, A. & Weinert, F. E. (1997). *Bedingungsfaktoren schulischer leistungen*. Max-Planck-Inst. für Psychologische Forschung.

Henderson, J. B., MacPherson, A., Osborne, J. & Wild, A. (2015). Beyond Construction: Five arguments for the role and value of critique in learning science. *International Journal of Science Education*, *37* (10), 1668–1697. doi:10.1080/09500693.2015.1043598

Hodson, D. (2014). Learning Science, Learning about Science, Doing Science: Different goals demand different learning methods. *International Journal of Science Education*, *36* (15), 2534–2553. doi:10.1080/09500693.2014.899722

Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary educational psychology*, *25* (4), 378–405. Elsevier.

Hofer, B. K. & Pintrich, P. R. (1997). The Development of Epistemological Theories: Beliefs About Knowledge and Knowing and Their Relation to Learning. *Review of Educational Research*, *67* (1), 88–140. doi:10.3102/00346543067001088

Hoffmann, R., Minkin, V. I. & Carpenter, B. K. (1997). Ockham's razor and chemistry. *International Journal for the Philosophy of Chemistry*, *3*, 3–28.

Hofstein, A. & Lunetta, V. N. (2004). The Laboratory in Science Education: Foundations for the Twenty-First Century. *Science Education*, *88* (1), 28–54. doi:10.1002/sce.10106

Hofstein, a. & Lunetta, V. N. (1982). The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research*, *52* (2), 201–217. doi:10.3102/00346543052002201

Holbrook, J. & Rannikmae, M. (2007). The Nature of Science Education for Enhancing Scientific Literacy. *International Journal of Science Education*, *29* (11), 1347–1362. doi:10.1080/09500690601007549

Höttecke, D. (2001). Die Vorstellungen von Schülern und Schülerinnen von der "Natur der Naturwissenschaften". *Zeitschrift für Didaktik der Naturwissenschaften*, *7*, 7–23.

Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6* (1), 1–55. doi:10.1080/10705519909540118

Inclusion Europe. (2016). *Informationen für alle: Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht Inclusion[Information for everyone: European rules on how to make information easy to read and understand]*. Europäischen Kommission. Verfügbar unter: www.life-long-learning.eu

Jamshidian, M. & Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, *75* (4), 649–674. Springer.

Jamshidian, M., Jalal, S. & Jansen, C. (2014). MissMech : An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). *Journal of Statistical Software*, *56* (6).

Johnstone, A. H. (1991). Thinking about thinking. *International Newsletter on Chemical Education*, *36*, 7–10.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance (Technical Report 37)*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Jones, T. & Sterling, D. R. (2011). Cooperative learning in an inclusive Classroom. *Sience scope*, *35* (3), 24–28.

Jorgensen, D., T., Pornprasertmanit, S., Schoemann, M., A. et al. (2018). *semTools: Useful tools for structural equation modeling*. Verfügbar unter: https://CRAN.R-project.org/package=semTools

Kahn, S. & Lewis, A. R. (2014). Survey on Teaching Science to K-12 Students with Disabilities: Teacher Preparedness and Attitudes. *Journal of Science Teacher Education*, *25* (8), 885–910. doi:10.1007/s10972-014-9406-z

Kampa, N., Neumann, I., Heitmann, P. & Kremer, K. (2016). Epistemological beliefs in science-a person-centered approach to investigate high school students' profiles. *Contemporary Educational Psychology*, *46*, 81–93. Elsevier Inc. doi:10.1016/j.cedpsych.2016.04.007

Khishfe, R. & Abd-El-Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry-oriented instruction on sixth graders' views of nature of science. *Journal of Research in Science Teaching*, *39* (7), 551–578. doi:10.1002/tea.10036

King-Sears, M. E., Johnson, T. M., Berkeley, S., Weiss, M. P., Peters-Burton, E. E., Evmenova, A. S. et al. (2015). An Exploratory Study of Universal Design for Teaching Chemistry to Students With and Without Disabilities. *Learning Disability Quarterly*, *38* (2), 84–96. doi:10.1177/0731948714564575

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Pozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54* (2), 222–237.

Kline, R. B. (2016). *Principles and Practice of Structural Equatuon Modeling*. New York: Guilford Press.

Koenig, J. A. & Bachman, L. F. (Hrsg.). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*. Washington, DC: National Academies Press.

Komorek, M. & Duit, R. (2004). The teaching experiment as a powerful method to develop and evaluate teaching and learning sequences in the domain of non-linear systems. *International Journal of Science Education*, *26* (5), 619–633. doi:10.1080/09500690310001614717

Kosecoff, J. & Klein, S. (1974). Instructional sensitivity statistics appropriate for objectives-based test items. In *Paper presented at the Annual Conference of the National Council on Measurement in Education*. Chicago, IL.

Krell, G. (2013). Vielfältige Perspektiven auf Diversity: erkunden, enthüllen, erzeugen. *Diversity ent-decken. Reichweiten und Grenzen von Diversity Policies an Hochschulen.*, 61–78. Weinheim: Beltz Juventa.

Kremer, K. (2010). *Die Natur der Naturwissenschaften verstehen: Untersuchungen zur Struktur und Entwicklung von Kompetenzen in der Sekundarstufe I[Undersatnding the Nature of Science: Investigations about structure and develpoment of competencies in secondary schools]* (S. 171). Universität Kassel.

Krippendorff, K. (2004). Reliability in Content Analysis : Some Common Misconceptions and Recommendations Reliability in Content Analysis : Some Common Misconceptions and. *Human Communication Research*, *30* (3), 411–433.

Kuckartz, U. (2012). *Qualitative lnhaltsanalyse. Methoden, Praxis, Computerunterstützung* (2. Auflage). Beltz Juventa. doi:10.1007/978-3-540-33306-7_5

Labudde, P. (2010). *Fachdidaktik Naturwissenschaft 1.-9. Schuljahr* (Band 3248). UTB.

Lamprianou, I. & Boyle, B. (2004). Accuracy of measurement in the context of mathematics National Curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement*, *41* (3), 239–259. doi:10.1111/j.1745-3984.2004.tb01164.x

Lawson, A. (1982). The nature of advanced reasoning and science instruction. *Journal of Research in Science Teaching*, *19*, 743–760.

Lederman, N. G. (1992). Students and Teachers Conceptions of the Nature of Science - a Review of the Research. *Journal of Research in Science Teaching*, *29* (4), 331–359. doi:10.1002/tea.3660290404

Lederman, N. G. (2013). Nature of science: Past, present, and future. In *Handbook of research on science education* (S. 845–894). Routledge.

Lederman, N. G. & AbdElKhalick, F. (1998). Avoiding De-Natured Science: Activities That Promote. In *The Nature of Science in Science Education: Rationales and Strategies* (S. 83–126). Netherlands: Kluwer Academic Publishers.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L. & Schwartz, R. S. (2002). Views of Nature of Science Questionnaire: Toward Valid and Meaningful Assessment of Learners' Conceptions of Nature of Science. *Journal of Research in Science Teaching*, *39* (6), 497–521. doi:10.1002/tea.10034

Lee, O., Miller, E. & Januszyk, R. (Hrsg.). (2015). *NGSS for all students*. Arlington: NSTA Press.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83* (404), 1198–1202. Taylor & Francis. doi:10.1080/01621459.1988.10478722

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guilford Press.

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R. & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22* (3), 486–506. doi:10.1037/met0000075

Lovett, B. J. & Lewandowski, L. J. (2015). *Testing Accomdations for Students with Disabilities*. Washington, DC: American Psychological Association.

Mahaffy, P. (2004). The Future Shape of Chemistry Education. *Chemistry Education Research and Practice*, *5* (3), 229. doi:10.1039/b4rp90026j

Markic, S. & Abels, S. (2014). Heterogeneity and Diversity: A Growing Challenge or Enrichment for Science Education in German Schools? *EURASIA Journal of Mathematics, Science & Technology Education*, *10* (4), 271–283. doi:10.12973/eurasia.2014.1082a

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47* (2), 149–174. Springer.

Mayring, P. (2013). Qualitative Inhaltsanaiyse. In *Qualitative Forschung: Ein Handbuch* (S. 468–475). Reinbeck bei Hamburg: Rowohlt. doi:10.1007/978-3-531-91182-3

McComas, W. F. & Olson, J. K. (1998). The Nature of Science in Science Education: Rationales and Strategies. In W. McComas (Hrsg.), *The nature of science in science education: Rationales andstrategies* (S. 41–52). Dordrecht: Kluwer Academic Publishers. doi:10.1017/CBO9781107415324.004

McGinnis, J. R. (2013). Teaching Science to Learners With Special Needs. *Theory Into Practice*, *52* (1), 43–50. doi:10.1080/07351690.2013.743776

Meade, A. W. & Lautenschlager, G. J. (2004). Same Question, Different Answers : CFA and Two IRT Approaches to Measurement Invariance. In *Symposium presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology*. Chicago, IL.

Menthe, J., Abels, S., Blumberg, E., Fromme, T., Marohn, A., Nehring, A. et al. (2017). Netzwerk inklusiver naturwissenschaftlicher Unterricht [Network and Work of Inclusive Science Education]. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 800–803). Regensburg: Gesellschaft fur Didaktik der Chemie und Physik (GDCP).

Menthe, J. & Hoffmann, T. (2015). Inklusiver Chemieunterricht: Chance und Herausforderung [Inclusive chemistry lessons: opportunity and challenge]. In J. Riegert & O. Musenberg (Hrsg.), *Inklusiver Fachunterricht in der Sekundarstufe* (S. 131–164). Stuttgart: Verlag W. Kohlhammer.

Menthe, J. & Sander, R. (2016). Mit Heterogenität umgehen. Sicheres Arbeiten im inklusiven und zieldifferenten Chemieunterricht. *Naturwissenschaften im Unterricht Chemie*, *156*, 45–46.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*, 215–237.

Milgram, D. (2011). How to recruit women and girls to the science, technology, engineering, and math (STEM) classroom. *Technology and engineering teacher*, *71* (3), 4–11. ERIC.

Mitchel, D. (2014). *What really works in Special and inclusive Education* (2. Auflage). Oxon, New York: Routledege.

Mulvey, B. K., Chiu, J. L., Ghosh, R. & Bell, R. L. (2016). Special education teachers' nature of science instructional experiences. *Journal of Research in Science Teaching*, *53* (4), n/a–n/a. doi:10.1002/tea.21311

National Research Council. (1998). *High stakes: testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press.

Naumann, A., Hartig, J. & Hochweber, J. (2017). Absolute and relative measures of instructional sensitivity. *Journal of Educational and Behavioral Statistics*, *Advance On* (X), 1–28. doi:10.3102/1076998617703649

Nehring, A., Nowak, K. H., Upmeier, A. & Tiemann, R. (2015). Predicting Students ' Skills in the Context of Scientific Inquiry with Cognitive, Motivational, and Sociodemographic Variables. *International Journal of Science Education*, *37* (9), 1343–1363. doi:10.1080/09500693.2015.1035358

Neumann, I. & Kremer, K. (2013). Nature of Science und epistemologische Überzeugungen – Ähnlichkeiten und Unterschiede. *Zeitschrift für die Didaktik der Naturwissenschaften*, *19*, 209–232.

Newsom, J. T. (2015). *Longitudinal Structural Equation Modeling*. New York: Routledge.

Newsom, J. T. (2018). Missing Data and Missing Data Estimation in SEM. In *Psy 523/623 Structural Equation Modeling* (S. 1–3).

NGSS Lead States. (2013). Next generation science standards: For states, by states. Washington, DC: National Academies Press.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R. & Duschl, R. (2003). What "Ideas-about-science" Should be taught in School Science? A Delphi study of the expert community. *Journal of Research in Science Teaching*, *40* (7), 692–720. doi:10.1002/tea.10105

Oser, F. & Blömeke, S. (2012). Überzeugungen von Lehrpersonen. Einführung in den Thementeil. *Zeitschrift für Pädagogik*, *58* (4), 415–421.

Patterson, A., Roman, D., Friend, M., Osborne, J. & Donovan, B. (2018). Reading for meaning: The foundational knowledge every teacher of science should have. *International Journal of Science Education*, *0* (0), 1–17. Taylor & Francis. doi:10.1080/09500693.2017.1416205

Phillips, S. E. (1994). High-Stakes Testing Accommodations: Validity Versus Disabled Rights. *Applied Measurement in Education*, *7* (2), 93–120. Routledge. doi:10.1207/s15324818ame0702_1

Pintrich, P. R. (2002). Future challenges and directions for theory and research on personal epistemology. *Personal epistemology: The psychology of beliefs about knowledge and knowing*, 389–414.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29* (4), 3–14. doi:10.1111/j.1745-3992.2010.00189.x

Praetorius, A. K. & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM - Mathematics Education*, *50* (3), 535–553. Springer Berlin Heidelberg. doi:10.1007/s11858-018-0946-0

Rao, K., Ok, M. W. & Bryant, B. R. (2014). A Review of Research on Universal Design Educational Models. *Remedial and Special Education*, *35* (3), 153–166. doi:10.1177/0741932513518980

Rapp, W. & Arndt, K. (2012). *Teaching Everyone: An Introduction to Inclusive Education*. Baltimore: Brookes Publishing. doi:10.1017/CBO9781107415324.004

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Verfügbar unter: https://www.r-project.org/

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. Verfügbar unter: https://CRAN.R-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É. & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17* (3), 354–373. doi:10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modelingle. *Journal of Statistical Software*, *48* (2), 1–36.

Rowe, M. (1974). A humanistic intent: The program of preservice elementary education at the University of Florida. *Science Education*, *58*, 369–376.

Salas-Wright, C. P., Olate, R. & Vaughn, M. G. (2013). Assessing empathy in Salvadoran high-risk and gang-involved adolescents and young adults: a Spanish validation of the basic empathy scale. *International Journal of Offender Therapy and Comparative Criminology*, *57* (11), 1393–1416. doi:10.1177/0306624X12455170

Salvia, J., Ysseldyke, J. E. & Witmer, S. (2017). What Test Scores means & UDA. In *Assessment in Special and Inclusive Education*. Cenage Learning.

Savolainen, H., Engelbrecht, P., Nel, M. & Malinen, O.-P. (2012). Understanding teachers' attitudes and self-efficacy in inclusive education: implications for pre-service and in-service teacher education. *European Journal of Special Needs Education*, *27* (1), 51–68. doi:10.1080/08856257.2011.613603

Schaefer, G. (1971). Fach – Didaktik – Fachdidaktik. Eine Standortbestimmung am Beispiel der Biologie. *MNU Journal*, *24* (7), 390–396.

Schimank, U. (2015). Grundriss einer integrativen Theorie der modernen Gesellschaft. *Zeitschrift für Theoretische Soziologie*, *4* (2), 236–268.

Schlüter, A.-K., Melle, I. & Wember, F. B. (2016). Unterrichtsgestaltung in Klassen des Gemeinsamen Lernens. *Sonderpädaoggische Förderung heute*, *61* (3), 270–285.

Schmidt, C. (2013). Analyse von Leitfadeninterviews. In *Qualitative Forschung: Ein Handbuch* (S. 447–456). Reinbeck bei Hamburg: Rowohlt.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of educational psychology*, *82* (3), 498. American Psychological Association.

Schraw, G., Bendixen, L. D. & Dunkle, M. E. (2002). Development and validation of the Epistemic Belief Inventory (EBI). Lawrence Erlbaum Associates Publishers.

Schwab, S. & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen [Verification of measurement invariance using CFA and DIF analyzes]. *Empirische Sonderpädagogik*, *3*, 175–193.

Schwartz, R. S., Lederman, N. G. & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education*, *88* (4), 610–645. doi:10.1002/sce.10128

Scruggs, T. E., Mastropieri, M. A., Berkeley, S. & Graetz, J. E. (2010). Do Special Education Interventions Improve Learning of Secondary Content? A Meta-Analysis. *Remedial and Special Education*, *31* (6), 437–449. doi:10.1177/0741932508327465

Scruggs, T. & Mastropieri, M. (2007). Science Learning in Special Education: The Case for Constructed Versus Instructed Learning. *Exceptionality*, *15* (2), 57–74. doi:10.1080/09362830701294144

Serenelli, F. & Mangiatordi, A. (2013). Universal design for learning: A meta-analytic review of 80 abstracts from peer reviewed journals. *Research on Education and Media*, *5* (1), 109–118.

Sharma, U. & Sokal, L. (2015). The impact of a teacher education course on pre-service teachers' beliefs about inclusion: An international comparison. *Journal of Research in Special Educational Needs*, *15* (4), 276–284. doi:10.1111/1471-3802.12043

Shepard, L., Taylor, G. & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's Grade 4 mathematics performance assessment*. os Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards,; Student Test.

Sideridis, G. D., Tsaousis, I. & Al-harbi, K. A. (2015). Multi-Population Invariance With Dichotomous Measures: Combining Multi-Group and MIMIC Methodologies in Evaluating the General Aptitude Test in the Arabic Language. *Journal of Psychoeducational Assessment*, *33* (6), 568–584. doi:10.1177/0734282914567871

Sireci, S. G., Banda, E. & Wells, C. S. (2018). Promoting Valid Assessment of Students with Disabilities and English Learners. In *Handbook of Accessible Instruction and Testing Practices* (S. 231–246). Springer.

Sireci, S. G., Scarpati, S. E. & Li, S. (2005). Test Accommodations for Students With Disabilities : An Analysis of the Interaction Hypothesis. *Review of Educational Research*, *75* (4), 457–490.

Sliwka, A. (2012). Diversität als Chance und als Ressource in der Gestaltung wirksamer Lernprozesse [Diversity as an opportunity and as a resource in designing effective learning processes]. In *Das interkulturelle Lehrerzimmer: Perspektiven neuer deutscher Lehrkräfte auf den Bildungs- und Integrationsdiskurs* (S. 169–176). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-94344-2_16

Specht, J., McGhie-Richmond, D., Loreman, T., Mirenda, P., Bennett, S., Gallagher, T. et al. (2016). Teaching in inclusive classrooms: efficacy and beliefs of Canadian preservice teachers. *International Journal of Inclusive Education*, *20* (1), 1–15. doi:10.1080/13603116.2015.1059501

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting Differential Item Functioning With Confirmatory Factor Analysis and Item Response Theory : Toward a Unified Strategy. *Journal of Applied Psychology*, *91* (6), 1292–1306. doi:10.1037/0021-9010.91.6.1292

Steele, M. (2008). Teaching strategies can help students with learning disabilities improve their performance in the science classroom. *The Science Teacher*, *75* (3), 38–42.

Stellungnahme der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE). (2017). *Inklusion: Bedeutung und Aufgabe für die Erziehungswissenschaft.* (S. 1–4). Stellungnahme der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE).

Stroh, M. (2014). Inklusion im naturwissenschaftlichen Unterricht – Beschreibung eines Spannungsfeldes. *Schulpädagogik heute*, *5* (9), 1–12.

The Center for Universal Design. (1997). The Principles of Universal Design (Version 2.0.). Raleigh, NC: North Carolina State University. Zugriff am 8.5.2018. Verfügbar unter: https://projects.ncsu.edu/ncsu/design/cud/about{_}ud/docs/use{_}guidelines.pdf

Therrien, W. J., Taylor, J. C., Hosp, J. L., Kaldenberg, E. R. & Gorsh, J. (2011). Science Instruction for Students with Learning Disabilities: A Meta-Analysis. *Learning Disabilities Research & Practice*, *26* (4), 188–203. doi:10.1111/j.1540-5826.2011.00340.x

Therrien, W. J., Taylor, J. C., Watt, S. & Kaldenberg, E. R. (2014). Science Instruction for Students With Emotional and Behavioral Disorders. *Remedial and Special Education*, *35* (1), 15–27. doi:10.1177/0741932513503557

Thompson, S. J., Johnstone, C. J. & Thurlow, M. L. (2002). *Universal design applied to large scale assessments (Synthesis Report 44)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., Thurlow, M. & Malouf, D. B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*, *6* (1), 1–15.

Thurlow, M., Christian, E. & Rogers, C. (2012). A Summary of the Research on the Effects of Test Accommodations: 2009-2010 (NCEO Technical Report). *University of Minnesota, Institute on Community Integration, National Center on Educational Outcomes (NCEO)*, *65*. University of Minnesota, Institute on Community Integration, National Center on Educational Outcomes (NCEO).

Thurlow, M., Lazarus, S. & Rogers, C. (2014). A Summary of the Research on the Effects of Test Accommodations, 2011-2012 (NCEO Synthesis Report). *University of Minnesota, Institute on Community Integration, National Center on Educational Outcomes (NCEO)*, *94*.

Tobin, R. & Tippett, C. D. (2013). Possibilities and Potential Barriers: Learning To Plan for Differentiated Instruction in Elementary Science. *International Journal of Science and Mathematics Education*, *12* (2), 423–443. doi:10.1007/s10763-013-9414-z

Torsheim, T., Cavallo, F., Levin, K. A., Schnohr, C., Mazur, J., Niclasen, B. et al. (2016). Psychometric Validation of the Revised Family Affluence Scale: a Latent Variable Approach. *Child Indicators Research*, *9* (3), 771–784. Child Indicators Research. doi:10.1007/s12187-015-9339-x

Trautmann, M. & Wischer, B. (2011). *Heterogenität in der Schule*. Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s13398-014-0173-7.2

Unger, B. (2018). Educationally Reconstructed Evolution Course - Evidence-Based Teaching in the Education of Future German Science Teachers. In *E-Book Proceedings of the ESERA 2017 Conference*. Dublin: Dublin City University (DCU).

Urban, D. & Mayerl, J. (2014). *Strukturgleichungsmodellierung. Ein Ratgeber für die Praxis*. Wiesbaden: Springer Fachmedien. doi:10.1017/CBO9781107415324.004

Urhahne, D., Kremer, K. & Mayer, J. (2008). Welches Verständnis haben Jugendliche von der Natur der Naturwissenschaften? Entwicklung und erste Schritte zur Validierung eines Fragebogens [What is adolescents' understanding of the nature of science? Development and first steps to validation of a ques.

Ve Wittig, M.-T. (2014). *Heterogenität - Belastung oder pädagogische Herausforderung? Eine Untersuchung von Lehrertypen an staatlichen Berliner Berufsschulen im Berufsfeld Wirtschaft und Verwaltung in Bezug auf den Umgang mit Schülervarianzen*. Dissertation. Humboldt Universität zu Berlin.

Villanueva, M. G. & Hand, B. (2011). Science for All: Engaging Students with Special Needs in and About Science. *Learning Disabilities Research & Practice*, *26* (4), 233–240. doi:10.1111/j.1540-5826.2011.00344.x

Villanueva, M. G., Taylor, J., Therrien, W. & Hand, B. (2012). Science education for students with special needs. *Studies in Science Education*, *48* (March 2015), 187–215. doi:10.1080/14703297.2012.737117

Walkowiak, M. & Nehring, A. (2017a). Eine inklusive Lernumgebung ist nicht genug: Fachspezifik, Theoretisierung und inklusive Unterrichtsentwicklung in den Naturwissenschaftsdidaktiken [An inclusive learning environment is not enough: theorization and inclusive teaching development in the sc. *Zeitschrift für Inklusion*, *03*, o.S.

Walkowiak, M. & Nehring, A. (2017b). Selbst und Fremdreflexion bei der Gestaltung und Umsetzung von inklusivem Chemieunterricht. In L. Schulze Heuling (Hrsg.), *How the Inclusive Classroorn Brings Fresh Ideas to Science and Education Inhaltsverzeichnis Introduction* (S. 77–83). Flensbu1g University Press.

Werning, R. & Baumert, J. (2013). Inklusion entwickeln: Leitideen für Schulentwicklung und Lehrerbildung. In *Inklusion: Forschungsergebnisse und Perspektiven* (S. 38–55). Oldenbourg.

Werning, R. & Thoms, S. (2017). Anmerkungen zur Entwicklung der schulischen Inklusion in Niedersachsen. *Zeitschrift für Inklusion, [S.l.], mai 2004*, *2*.

Weston, T. J. (2002). *The validity of oral accommodation in testing (NCES 200306)*. Washington, DC: National Center for Education Statistics.

Wilde, M. (2018). Question Pro. Zugriff am 20.6.2018. Verfügbar unter: https://www.questionpro.com/blog/de/

Williamson Sprague, E. & Dahl, D. W. (2010). Learning to click: An evaluation of the personal response system clicker technology in introductory marketing courses. *Journal of Marketing Education*, *32* (1), 93–103. SAGE Publications Sage CA: Los Angeles, CA.

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

Winter, P. C., Kopriva, R. J., Chen, C. S. & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, *16* (4), 267–276. doi:10.1016/j.lindif.2007.01.001

Wise, S. L. & Kingsbury, G. G. (2016). Modeling Student Test-Taking Motivation in the Context of an Adaptive Achievement Test. *Journal of Educational Measurement*, *53* (1), 86–105. doi:10.1111/jedm.12102

Witmer, S., Schmitt, H., Clinton, M. & Mathes, N. (2018). Accommodation Use During Content Area Instruction for Students with Reading Difficulties: Teacher and Student Perspectives. *Reading & Writing Quarterly*, *34* (2), 174–186. Taylor & Francis.

Wocken, H. (2014). Frei herumlaufende Irrtümer. Eine Warnung vor pseudoinklusiven Betörungen. *Gemeinsam Leben*, *1*, 52–54.

Wu, A. D., Liu, Y., Gadermann, A. M. & Zumbo, B. D. (2010). Multiple-indicator multilevel growth model: A solution to multiple methodological challenges in longitudinal studies. *Social Indicators Research*, *97* (2), 123–142. doi:10.1007/s11205-009-9496-8

Wu, A. D., Zhen, L. & Zumbo, B. D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis : A Demonstration With TIMSS Data. *Practical Assessment, Research & Evaluation.*, *12* (3), 1–26.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with {R} Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. Verfügbar unter: https://github.com/rstudio/bookdown

Xie, Y. (2018). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. Verfügbar unter: https://yihui.name/knitr/

Zhu, H. (2018). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. Verfügbar unter: https://cran.r-project.org/web/packages/kableExtra/index.html

9 | Danksagung und Lebenslauf

9.1 Danksagung

Eine Dissertation gleicht einem Marathon. Ohne Menschen, die diesen begleiten, ist der Lauf nicht zu schaffen. Ich danke daher Herrn Prof. Dr. Andreas Nehring. Andreas, Du hast mich seit meinem Studium mit deinem Engagement für *Guten Chemieunterricht* motiviert. Deine Perspektive auf die Dinge hat das ein ums andere Mal meine Arbeit bereichert. Ich danke in diesem Zusammenhang auch Herrn Prof. Dr. Sascha Schanze und Herrn Prof. Dr. Harald Groupengießer. Sascha und Harald, vielen Dank, dass Ihr die Ressourcen am IDN auch mir zu kommen lassen habt. Eure kritischen Fragen haben diese Arbeit inhaltlich befruchtet. Ich danke an dieser Stelle auch ganz herzlich Frau Prof. Insa Melle, die sich einerseits bereit erklärt hat, diese Dissertation mit zu begutachten. Außerdem haben Ihren Mitarbeiterinnen und Mitarbeiter Annka, Daggy und Thomas diese Arbeit von Anfang an positiv beeinflusst. Ich danke auch Frau Prof. Dr. Nadja Bigall für Ihre Unterstützung in der Prüfungskommission. Schließlich danke ich auch Herrn Prof. Dr. Dirk Lange, der diese Arbeit durch seine Bemühungen finanziell ermöglicht hat.

Ich bedanke mich bei allen Mitarbeitenden des IDN. Besonders danke ich Theresa, Robert, Nina, Niklas und Julian. Es war eine wunderbare Zeit mit euch, den Nachwuchs am IDN zu stellen. Ihr alle seid Unikate! Barnd, deine tägliche Lektion zur Verbesserung meines biologischen Wissens werde ich vermissen; ebenso deine Anfragen zur Bedeutung meiner statistischen Modelle! Dir allseits 'Gut Grün'!

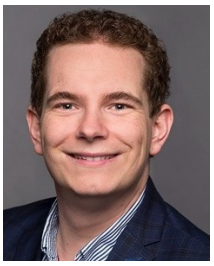
Ich danke weiterhin allen Lehrerinnen und Lehrern, die durch ihre Hilfe die Studien möglich gemacht haben. Ich danke allen Eltern und Schülerinnen und Schülern, die sich diesem Projekt geöffnet haben. Schließlich danke ich der R-Crowd, die durch ihre phänomenale Arbeit die Analysen und Abfassung dieser Arbeit ermöglicht haben. Ihr teilt großzügig Eure Zeit und Euer Wissen mit den Menschen. Außerdem danke ich den Hilfskräften Jesco, Leonardo, Claudia und Benjamin für ihre Mithilfe bei der Durchführung beider Studien.

Ich danke Mara. Ohne Deine Unterstützung wäre diese Arbeit nie entstanden. Deine Art hat die vielen kritischen Momente aufgefangen. Vermutlich habe ich das nicht immer angemessen würdigen können. Ich danke auch meinen Kindern. Leevi und Matilda, ihr musstet oft auf mich verzichten und habt nie deswegen geklagt. Vielmehr wart ihr immer interessiert an meiner Arbeit und *"der Uni"*. Ich danke meinen Eltern, die mich in der Schulzeit auf Spur gehalten und die Basis für mein heutiges Tun gelegt haben. Ich danke aber auch meiner ganzen Familie, die meine Umwege stets begleitet und unterstützt hat.

9.2 Lebenslauf

MALTE WALKOWIAK, M.Ed.
malte.walkowiak@gmx.de

PERSÖNLICHE DATEN



Geboren am 05.12.1988 in Hannover

STUDIUM

09/2015 bis heute	Promotionsstudium an der Naturwissenschaftlichen Fakultät, ***Leibniz Universität Hannover***
10/2013-09/2015	**Master of Education** (Lehramt) in Chemie und Philosophie/Ethik, ***Universität Potsdam,*** Abschluss 2015
04/2012-09/2013	**Bachelor of Education** (Lehramt) in Chemie und Philosophie/Ethik, ***Universität Potsdam,*** Abschluss 2013
10/2009-03/2012	**Bachelor of Education** (Lehramt) in Chemie und Ev. Religion, ***Humboldt Universität zu Berlin***
08/2005-07/2008	**Abitur**, ***Humboldtschule Hannover***

BERUFSERFAHRUNG

09/2018 bis heute	**AWS Cloud Consultant, *tecRacer Consulting GmbH Hannover,*** Machine Learning und Big Data-Infrastrukturen
09/2015 - 09/2018	**Wissenschaftlicher Mitarbeiter, *Leibniz Universität Hannover***, Institut für Didaktik der Naturwissenschaften

SCHLÜSSELQUALIFIKATIONEN

10/2018	**AWS Certified Solutions Architect**
07/2018	**Data Mining in R**
02/2018	**Betriebswirtschaftliche Grundlagen für Naturwissenschaftler**
IT	**C, Python, R und Markdown**