Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Vergleichsarbeiten (VERA): Eine empirische Untersuchung zur Akzeptanz bei Lehrkräften

Vom Promotionsausschuss des Fachbereichs 8: Psychologie der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. phil.) genehmigte Dissertation

von

Johanna Detzel

Vorsitzende des Promotionsausschusses: Prof. Dr. Tanja Lischetzke
Erstgutachter: Prof. Dr. Ingmar Hosenfeld
Zweitgutachterin: Vertr.-Prof. Dr. habil. Julia Riebel

Datum der wissenschaftlichen Aussprache: 09. August 2024

Inhaltsverzeichnis

Zusammenfassung

Die Forschungsliteratur zu Vergleichsarbeiten (VERA) thematisiert seit ihrer Einführung die Akzeptanz von VERA bei Lehrkräften als eine entscheidende Gelingensbedingung einer erfolgreichen Implementierung des Instruments zur Schul- und Unterrichtsentwicklung. Diesen Untersuchungen mangelt es jedoch i. d. R. an einer theoretischen Grundlage sowie einer eindeutigen oder gar einheitlichen Konzeptualisierung und Begriffsdefinition. Anknüpfend an dieses Forschungsdesiderat beschäftigt sich diese Arbeit mit der Entwicklung einer theorie- und empiriegestützen Definition von Akzeptanz (Ziel 1) und der Konkretisierung dieses Akzeptanzverständnisses im Kontext von VERA, einschließlich der Entwicklung (Ziel 2) und Validierung (Ziel 3) eines empirischen Akzeptanzmodells. Des Weiteren widmet sich die Arbeit der Analyse von Unterschieden in der Akzeptanz zwischen Grundschullehrkräften (VERA3) und Lehrkräften weiterführender Schulen (VERA8) (Ziel 4). Theoretische Grundlage zur Bearbeitung dieser Forschungsziele bilden Einstellungstheorien und Forschungsbefunde zur Wahrnehmung von Lehrkräften mit Blick auf Vergleichsarbeiten. Basierend auf dem *Technology Acceptance Model (TAM)*, in Verbindung mit einschlägiger Forschungsliteratur zu Vergleichsarbeiten, werden eine Definition von Akzeptanz sowie ein entsprechendes Forschungsmodell aufgestellt. Zentrale Einflussfaktoren im Modell auf die Nutzungsintention der Rückmeldungen aus Vergleichsarbeiten stellen dabei die Wahrnehmung der zeitlichen Belastung und der Nützlichkeit dar, deren Einfluss auf Einstellung und Nutzungsintention untersucht wird. Die Überprüfung dieses Akzeptanzmodells erfolgt mittels Strukturgleichungsanalysen. Datengrundlage bilden Survey-Daten aus Befragungen mit Lehrkräften zu VERA3 (2018, N = 4 141; 2019, N = 2 751) und VERA8 (2019, N = 782). In der empirischen Überprüfung erweist sich das ursprünglich aufgestellte Modell 1 zwar als passend zu den Daten, wird jedoch aufgrund statistischer und inhaltlicher Überlegungen angepasst. Die Ergebnisse der empirischen Analyse des angepassten Modells 2, ohne eine separate Modellierung eines Aufwand-Nutzen-Konstrukts, zeigen hingegen, dass es gelungen ist, ein Forschungsmodell zur Akzeptanz von VERA aufzustellen, das sich in verschiedenen Stichproben (VERA3 und VERA8) bewährt. Zentrale Erkenntnis der Arbeit ist, dass die Nutzenwahrnehmung den wichtigsten Einflussfaktor bei der Frage nach der Weiterarbeit mit den Rückmeldungen aus Vergleichsarbeiten darstellt. Die Arbeit kann die Überlegenheit dieses Einflusses gegenüber einer zeitlichen Entlastung von Lehrkräften quantitativ belegen und verdeutlicht hierbei, dass eine Maximierung des Nutzens gegenüber einer zeitlichen Optimierung in der Praxis priorisiert werden sollte. Die Implikationen dieser Erkenntnis für Forschung und Praxis werden im letzten Teil der Arbeit diskutiert.

1. Einleitung

Anfang der 2000er wurden als Antwort auf den viel zitierten „PISA-Schock“, also das schlechte Abschneiden deutscher Schüler*innen in internationalen Vergleichsstudien, in der deutschen Bildungspolitik umfassende Reformmaßnahmen, wie die Einführung bundesweit geltender Bildungsstandards, beschlossen (Groß Ophoff, 2013; KMK, 2001, 2004a; Maaz, Emmrich, Kropf & Gärtner, 2019). 20 Jahre später erzielten deutsche Schüler*innen in der aktuellen PISA Studie 2022 in den Bereichen Lesekompetenz und Mathematik nach zwischenzeitlicher Verbesserung ein schlechteres Ergebnis als bei der ersten PISA Studie im Jahr 2000 (OECD, 2023).

Vor diesem Hintergrund darf die Effektivität der einzelnen Bestandteile der Reformen der Bildungspolitik in Frage gestellt werden, so unter anderem auch die Einführung bundesweiter Vergleichsarbeiten (VERA), deren Nutzen viele Lehrkräfte von Anfang an bezweifeln. Obgleich bereits vor zwei Jahrzehnten in den meisten Bundesländern eingeführt und mittlerweile in der gesamten Bundesrepublik[1] etablierter Bestandteil der jährlichen Schulroutine, sind Vergleichsarbeiten (VERA) auch heute noch ein umstrittenes Instrument in der deutschen Bildungslandschaft (Altrichter, Moosbrugger & Zuber, 2016; Wagner & Koch, 2021; Zimmer-Müller & Hosenfeld, 2013).

Während Leistungsvergleiche wie PISA, TIMSS oder PIRLS auf ein Systemmonitoring ausgelegt sind, sollten nationale Vergleichsarbeiten im Zuge der Implementierung dieser Bildungsstandards als Instrument der Schul- und Unterrichtsentwicklung dienen. Im Rahmen der jährlichen, bundesweit in den dritten (VERA3) und achten (VERA8) Klassen durchgeführten Testungen erhalten Lehrkräfte unmittelbare Rückmeldungen zu den Leistungen ihrer Schüler*innen. Auf dem Weg zur Erreichung des Bildungsstandards dienen diese als Hilfsmittel zur Überprüfung des Lernstandes ihrer Schüler*innen und liefern Impulse zur Weiterarbeit mit den Ergebnissen und zur Förderung von Schüler*innen, mit dem Ziel, perspektivisch eine Entwicklung und Verbesserung des Unterrichts zu erreichen (Altrichter et al., 2016; Diemer, Hartung-Beck & Kuper, 2013; Maier, Metz, Bohl, Kleinknecht & Schymala, 2012; F. Thiel, Tarkian, Lankes, Maritzen & Riecke-Baulecke, 2019).

[1] Das Bundesland Niedersachsen nimmt seit dem Schuljahr 2019/2020 nicht mehr an den Vergleichsarbeiten teil.

Auf Unterrichtsebene sind Lehrkräfte somit die zentralen Akteur*innen zur Implementation von Verbesserungen in Folge der Vergleichsarbeiten (Jäger, 2011). Erst durch eine aktive Auseinandersetzung mit den rückgemeldeten Daten und deren Nutzung können darauf basierende schulische Lehr-Lern-Prozesse gestaltet und weiterentwickelt werden und Vergleichsarbeiten ihre Wirkung einer datenbasierten Schul- und Unterrichtsentwicklung entfalten (Diemer & Kuper, 2011; Hartung-Beck & Diemer, 2009). In der Schulpraxis stoßen die Vergleichsarbeiten jedoch auf wenig positive Resonanz bei Lehrkräften, und die Nutzung der Ergebnisse bleibt deutlich hinter den Erwartungen zurück (siehe bspw. Maier, 2008a; Skejic, Neumann & Mangal, 2015; Wurster, Richter & Lenski, 2017). Dies zeigt nicht nur der Blick in die entsprechende Forschungsliteratur, es spiegelt sich auch in Aussagen von Lehrkräften, bspw. im Rahmen von Evaluationsbefragungen, wider. Zitate wie:

„Mich ärgert es aber auch enorm, dass im Laufe des 3. Schuljahres Wissen von Ende des Schuljahres abgetestet wird und Ihnen das bekannt ist. Das ist überhaupt nicht fair!!!“

„Es ist ja einfach Pflicht also müssen wir da durch. Manchmal ist es Mist, Beamter zu sein. Diese produzierten Überstunden sind einfach mit drin und werden von uns verlangt."

„Für die Entwicklung des eigenen Unterrichts völlig unbrauchbar, da nicht die Kompetenzen getestet werden die auf den Heften drauf stehen.“

(Lehrkräftebefragung VERA3, zepf, 2019[2]) benennen nur einige exemplarische Kritikpunkte, die seit Jahren von Lehrkräften mit Blick auf Vergleichsarbeiten geäußert werden.

In der einschlägigen Forschungsliteratur, die sich mit Problemen von Vergleichsarbeiten in der Praxis beschäftigt, wird eine fehlende Akzeptanz auf Seiten schulischer Akteur*innen häufig thematisiert. Diese wird oftmals in direkter Verbindung mit der Nutzung von Ergebnisrückmeldungen und der damit einhergehenden Weiterarbeit mit VERA untersucht. Die Akzeptanz des Verfahrens gilt hierbei als entscheidende Voraussetzung einer auf VERA basierten Schul- und Unterrichtsentwicklung (siehe bspw. Groß Ophoff, Koch & Hosenfeld, 2019; Kühle & Peek, 2007; Maier, 2008a, 2009c; 2010a; Vogel, Blum, Achmetli & Krawitz, 2016; Vogel, 2020; Wagner, Hosenfeld & Zimmer-Müller, 2019). In keiner dieser Arbeiten aus dem Kosmos der Vergleichsarbeiten wird jedoch der Begriff Akzeptanz definiert. Die inhaltliche Bedeutung

[2] Die aufgeführten Zitate stammen aus bisher unveröffentlichten Lehrkräftebefragungen im Rahmen der jährlich vom Zentrum für Empirische Pädagogische Forschung (zepf) im Zuge der VERA-Testung durchgeführten Evaluationsbefragungen.

ergibt sich lediglich aus dem Kontext bzw. in quantitativen Arbeiten aus der Operationalisierung oder dem Funktionsverständnis von Akzeptanz mit Blick auf andere Faktoren in der Wahrnehmung von oder im Umgang mit Vergleichsarbeiten. Eine differenzierte Begriffsbestimmung erfolgt nicht. Insgesamt lässt sich aus der Literatur in Bezug auf VERA ein eher undifferenziertes Verständnis von Akzeptanz als eine Art positiv konnotierte Voraussetzung für die Rezeption und Nutzung von VERA-Rückmeldungen ableiten. Häufig finden sich dabei inhaltliche Überschneidungen zu Konzepten der Einstellungsforschung (siehe bspw. Diemer, 2013; Maier, 2008a, 2008b; Wagner et al., 2019).

Im Bereich der Einstellungsforschung lassen sich auch die Ursprünge von Akzeptanzforschung verorten. Obgleich in der pädagogischen Psychologie generell keine eindeutige Begriffsdefinition von Akzeptanz ausgemacht werden kann, umspannt dieses Konstrukt in verwandten Forschungsdisziplinen wie Sozialpsychologie oder Konsumentenforschung ein etabliertes Forschungsfeld (Lucke, 1995). Insbesondere aus der Einstellungsforschung hervorgegangene Untersuchungen zur Akzeptanz technologischer Innovationen, wie die Arbeiten von F. D. Davis, Bagozzi und Warshaw (1989), F. D. Davis (1993) und Venkatesh, Morris, Davis und Davis (2003) zum *Technology Acceptance Model* (TAM), prägen hierbei den Begriff. Dabei sind Konstrukte wie Einstellung oder Nützlichkeit von Bedeutung, die auch in der Literatur zur Akzeptanz von Vergleichsarbeiten häufig gebraucht werden (siehe bspw. Diemer, 2013; Maier, 2008a, 2008b; Wagner et al., 2019; Wurster, Bach et al., 2016), dort jedoch ohne klare Struktur und Konzeptualisierung.

1.1. Fragestellung und Zielsetzung

Aus dieser aufgedeckten Forschungslücke einer fehlenden Konzeption bzw. Definition von Akzeptanz im Kontext von VERA zusammen mit der praktischen Relevanz ungenutzter Potenziale von Vergleichsarbeiten infolge einer häufig ablehnenden Haltung von Lehrkräften, leitet sich die übergeordnete Zielsetzung der vorliegenden Arbeit ab: Das Ziel dieser Arbeit ist es, eine theoretisch fundierte Durchdringung des Begriffs Akzeptanz, bezogen auf Vergleichsarbeiten. Auf dieser Basis soll das Verständnis für die Wahrnehmung und das Verhalten von Lehrkräften im Hinblick auf VERA verbessert und Ansätze für eine Optimierung und Weiterentwicklung des Verfahrens identifiziert werden.

Das übergeordnete Forschungsziel beinhaltet sowohl eine theoretische als auch eine empirische Fragestellung. Diese lassen sich jeweils in zwei Forschungsziele differenzieren. Forschungsziel 1 und 2 zielen auf eine eher theoretische und konzeptionelle Zielsetzung ab, Forschungsziel 3 und 4 stellen dagegen klassische empirische Fragestellungen dar:

Ziel 1: Herleitung einer theorie- und empiriegeleiteten Definition von Akzeptanz.
Dieses zentrale Erkenntnisinteresse zielt auf die Begriffsklärung und theoretische Einbettung des Konstrukts Akzeptanz ab. Einstellungstheorien wie die *Theory of Reasoned Action (*TRA), die *Theory of Planned Behavior* (TPB) und das *Technology Acceptance Model* (TAM) werden herangezogen, um eine theoretische Grundlage für die Akzeptanzuntersuchung im Kontext von VERA und eine entsprechende Akzeptanzdefinition zu schaffen.

Ziel 2: Konkretisierung des Akzeptanzverständnisses im Kontext von VERA und Konzeption eines empirisch überprüfbaren Modells.
Neben der Erarbeitung einer theoriebasierten Definition von Akzeptanz ist es ein Anliegen dieser Arbeit, dieses Akzeptanzverständnis im Kontext von Vergleichsarbeiten weiter auszudifferenzieren. Basierend auf einschlägiger Forschungsliteratur sollen hierfür Wahrnehmungsaspekte von Lehrkräften mit Blick auf Vergleichsarbeiten und insbesondere konkrete Ursachen für die Nutzung oder Nicht-Nutzung von VERA-Ergebnissen identifiziert und in ein empirisch überprüfbares Modell übertragen werden.

Ziel 3: Empirische Validierung des aufgestellten Akzeptanzmodells.
Die Überprüfung der empirischen Bewährung des erarbeiteten Akzeptanzmodells im Kontext von VERA stellt ein weiteres Ziel dieser Arbeit dar. Hierzu werden anhand von Survey-Daten aus Lehrkräftebefragungen die im Modell postulierten Zusammenhänge mit Strukturgleichungsanalysen untersucht.

Ziel 4: Analyse von Unterschieden in der Akzeptanz bei Lehrkräften verschiedener Schularten.
Einzelne Forschungsergebnisse (siehe bswp. Maier & Rauin, 2006) und die Projektpraxis rund um die Durchführung von Vergleichsarbeiten geben Hinweise auf Unterschiede in der Wahrnehmung und Bewertung bei Lehrkräften unterschiedlicher Schultypen (Primarstufe vs. Sekundarstufe I). Daher liegt das letzte Forschungsziel dieser Arbeit in der Analyse potenzieller Unterschiede der Akzeptanz von Vergleichsarbeiten zwischen Lehrkräften verschiedener Schulformen, genauer zwischen Grundschullehrkräften (VERA3) und Lehrkräften weiterführender

Schulen (VERA8). Zur Untersuchung dieser Unterschiede werden im empirischen Teil dieser Arbeit latente Gruppenmodelle geschätzt.

1.2. Aufbau der Arbeit

Zur Bearbeitung der aufgestellten Forschungsziele folgt die Arbeit dem beschriebenen Aufbau: Nach der in diesem Kapitel vorgestellten Ausgangslage der Untersuchung und einer Darlegung der Relevanz der untersuchten Thematik für Forschung und Praxis wurden auf dieser Grundlage das Forschungsdesiderat und die zentralen Fragestellungen der Arbeit vorgestellt, und das gewählte Forschungsdesign knapp umrissen.

Der Theorieteil dieser Arbeit (Kapitel 2) gliedert sich in zwei Teile: der erste Teil (Kapitel 2.1) befasst sich mit der Aufarbeitung des Akzeptanzbegriffs, basierend auf Erkenntnissen der Einstellungsforschung. Hierbei steht die Darstellung verschiedener Einstellungstheorien, mit Fokus auf die Beziehung von Einstellung und Verhalten, sowie die damit einhergehende Herleitung einer für diese Arbeit gültigen Akzeptanzdefinition im Mittelpunkt. Das zweite Teilkapitel (Kapitel 2.2) des Theorieteils gibt zunächst einen Überblick über die Entstehung, Konzeption und Ziele der Vergleichsarbeiten (VERA) (Kapitel 2.2.1). Dem schließt sich eine Analyse des Forschungsstandes zu der Wahrnehmung von Lehrkräften im Hinblick auf Vergleichsarbeiten an, wobei ebenfalls der allgemeine Blick auf Bildungsstandards kurz thematisiert wird (Kapitel 2.2.2). Die Auseinandersetzungen zu Theorie und Forschungsstand zielen auf die Ausarbeitung des in Kapitel 3 dargestellten konzeptionellen Modells der Akzeptanz von Vergleichsarbeiten ab, aus dem die empirisch zu prüfenden Hypothesen abgeleitet werden. Im Rahmen der Erarbeitung dieses Forschungsmodells wird auch die dieser Arbeit zugrundeliegende Definition von Akzeptanz entwickelt.

Im sich anschließenden Kapitel 4 wird das gewählte methodische Vorgehen zur Überprüfung des in Kapitel 3 erarbeiteten empirischen Modells vorgestellt. Zunächst wird das Untersuchungsdesign (Kapitel 4.1) einschließlich des Erhebungsinstruments, der Erhebungsmethode und Durchführung der Datenerhebung, sowie das Vorgehen bei der Datenauswertung erläutert. In Kapitel 4.2 werden das Vorgehen der Datenaufbereitung und abschließend die finalen Stichproben beschrieben. Zum Abschluss des Methodenteils werden in Kapitel 4.3 die genutzten Analysemethoden erläutert. Im ersten Abschnitt wird die Auswahl der Methode der Strukturgleichungsmodellierung begründet (Kapitel 4.3.1), bevor im zweiten (Kapitel 4.3.2) und

dritten (Kapitel 4.3.3) Teilkapitel die Grundlagen von Strukturgleichungsmodellen sowie das gewählte Vorgehen vermittelt werden. Kapitel 4.3.4 beschäftigt sich zuletzt mit dem Konzept latenter Gruppenanalysen.

Kapitel 5 wendet sich der Darstellung der empirischen Ergebnisse zu. Hierbei werden zunächst die Ergebnisse der Analyse des in Kapitel 3 aufgestellten Forschungsmodells berichtet (Kapitel 5.1), bevor in Kapitel 5.2 eine notwendige Modellanpassung beschrieben wird. In Kapitel 5.3 erfolgt eine Validierung dieses angepassten finalen Modells. Das letzte Teilkapitel der empirischen Analyse widmet sich der Untersuchung latenter Gruppenunterschiede zwischen VERA3- und VERA8-Lehrkräften (Kapitel 5.4).

In Kapitel 6 erfolgt die kritische Zusammenfassung und Interpretation der gefundenen Ergebnisse vor dem Hintergrund der zuvor aufgestellten Forschungsziele (Kapitel 6.1), gefolgt von einer Auseinandersetzung mit den Limitationen der Untersuchung (Kapitel 6.2).Die Arbeit schließt mit Implikationen für die Umsetzung von Vergleichsarbeiten in der Praxis (Kapitel 6.3) und erläutert im finalen Kapitel 6.4 Implikationen für die Forschung und gibt einen Ausblick auf weitere mögliche zu untersuchende Anschlussfragestellungen.

2. Theoretische und thematische Einordnung

Die folgenden Kapitel legen den theoretischen und konzeptionellen Grundstein dieser Arbeit. Um das Konzept Akzeptanz aus einem theoretischen Blickwinkel zu beleuchten, widmet sich Kapitel 2.1 der Einstellungsforschung. Hierbei werden zunächst in Kapitel 2.1.1 die Grundlagen der Einstellungsforschung vermittelt, um ein solides Verständnis für die in Kapitel 2.1.2 dargestellten Einstellungstheorien zu schaffen, welche die Basis für die Konzeption des Akzeptanzbegriffs im Rahmen dieser Arbeit legen. Für ein erleichtertes Verständnis werden die erläuterten Theorien dabei mit fiktiven Beispielen aus dem Schulkontext illustriert. Zur thematischen Einordnung in den Kontext der Vergleichsarbeiten (VERA) richtet Kapitel 2.2 das Augenmerk auf die Entstehung, Konzeption und Ziele von Vergleichsarbeiten (Kapitel 2.2.1), gefolgt von einer Analyse des Forschungsstandes zur Akzeptanz und allgemeinen Wahrnehmung von Lehrkräften im Hinblick auf Bildungsstandards und Vergleichsarbeiten (Kapitel 2.2.2).

2.1. Akzeptanz in der Psychologie – von der Einstellung zur Akzeptanz

2.1.1. Grundlagen der Einstellungsforschung

2.1.1.1. Historische Entwicklung und Definitionen des Einstellungsbegriffs

Das Konstrukt Einstellung, aus dem im Englischen verwendeten Terminus *attitude* hervorgegangen, beschäftigt die sozialpsychologische Forschung bereits seit über einem Jahrhundert (Eckardt, 2015). Mit dem Aufkommen der Sozialpsychologie etwa zur Jahrhundertwende des 20. Jahrhunderts entstand ein zunehmendes Interesse an der Untersuchung mentaler Konstrukte. Die Ursprünge der Einstellungsforschung als sozialpsychologisches Forschungsfeld lassen sich auf die 20er und 30er Jahre des 20. Jahrhunderts zurückführen (Maio & Haddock, 2010; McGuire, 1986). Generell existieren zu kaum einem anderen Konstrukt der Sozialpsychologie derart zahlreiche theoretische und empirische Untersuchungen (Allport, 1967). Jedoch beschränkt sich Einstellungsforschung nicht auf den Bereich der Sozialpsychologie, sondern erstreckt sich auch in andere Disziplinen, wie Soziologie, Verhaltenspsychologie, Konsumentenforschung sowie in Bereiche der Politikwissenschaft (Maio & Haddock, 2010). Die

Einstellungsforschung befasst sich dabei mit der Untersuchung verschiedenster Konstrukte, wie Wahrnehmung, Kognitionen oder Emotionen (Eagly & Chaiken, 2005).

Trotz der intensiven, auch interdisziplinären Forschung und einer Vielzahl an wissenschaftlichen Erkenntnissen handelt es sich bei dem Konstrukt Einstellung, wie die folgenden Abschnitte zeigen werden, keineswegs um einen einfach zu definierenden Begriff, da in der historischen Entwicklung eine Vielzahl an Ansätzen verfolgt wurden: Grundsteine der Einstellungsforschung legten u. a. Thurstone (1928) mit der Entwicklung einer Skala zur Einstellungsmessung und Likert (1932) mit der Einführung der bis heute gebräuchlichen Likert-Skala. Als einer der ersten Forscher, die sich intensiv mit dem Einstellungskonzept befassten, beschreibt Thurstone (1928, 1931) Einstellungen als „the sum total of a man's inclinations and feelings, prejudice or bias, pre- conceived notions, ideas, fears, threats, and convictions about any specified topic" (Thurstone, 1928, S. 531). In seiner Arbeit betrachtet er Einstellungen als eindimensionales Konstrukt und definiert diese als „the affect for or against a psychological object" (Thurstone, 1931, S. 261). Eine Einstellung ist dabei immer subjektiv und individuell (Thurstone, 1928). Diese Definition stellt die Wichtigkeit einer einzelnen Komponente – des Affekts (*affect*)– zur Repräsentation einer Einstellung heraus und kann daher als ein Einkomponentenmodell der Einstellung angesehen werden. Der Begriff Affekt bezieht sich im Kontext der Einstellungsforschung auf Empfindungen und Emotionen (Agarwal & Malhotra, 2005; Malhotra, 2005). Gemäß der Definition von Thurstone (1931) lässt sich Einstellung demnach als die (gefühlsmäßige) Bewertung eines psychologischen Objektes im Sinne einer Zustimmung oder Ablehnung interpretieren. Einstellungen sind dabei auf einem Kontinuum von positiv zu negativ bzw. von vorteilhaft zu unvorteilhaft angesiedelt (Ajzen & Fishbein, 1980). Nicht notwendigerweise muss der Ausdruck einer bestimmten Einstellung dabei auch in einem damit korrespondierenden Verhalten resultieren (Thurstone, 1928).

Allport (1935, 1967) dagegen hebt in seiner bis in die 60er Jahre des letzten Jahrhunderts prägenden Definition die Beziehung von Einstellung und Verhalten hervor und beschreibt Einstellung darin als „a mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related" (Allport, 1967, S. 6–7). Die Definition impliziert zum einen, dass Einstellungen zunächst auf Erfahrung basieren und sich zum anderen auf das Verhalten einer Person auswirken. Die verhaltensleitende Eigenschaft von Einstellungen stand lange im Zentrum der Einstellungsforschung und ist häufig in Definitionen – insbesondere in frühen – zu finden (Fazio, 1986). Diese weitverbreitete Annahme einer Beziehung von Einstellung und Verhalten

ist in der wissenschaftlichen Debatte jedoch nicht unumstritten. Speziell im Laufe der 60er Jahre stellen verschiedene Untersuchungen die Einstellungs-Verhaltens-Relation in Frage und kommen zu dem Schluss, dass Einstellungen nicht zwangsläufig als Verhaltensprädiktoren angesehen werden können (Festinger, 1964; Wicker, 1969).

Einem differenzierteren Definitionsansatz folgt Katz (1960), der in seiner Arbeit drei Facetten von Einstellungen unterscheidet: affektive, kognitive und verhaltensbezogene Einstellungskomponenten. Generell folgen neuere Ansätze der Einstellungsforschung häufig dieser Dreiteilung (Haddock & Maio, 2014, 2019). Auch in dem vielzitierten Standardwerk „The psychology of attitudes“ von Eagly und Chaiken (1993) ist diese Unterscheidung zu finden. Die Autorinnen definieren Einstellung als „psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor“ (Eagly & Chaiken, 1993, S. 1). Einstellungen sind demnach als summarische Bewertungen bestimmter Sachverhalte oder Objekte zu verstehen. Die psychologische Tendenz bezieht sich dabei auf den inneren Zustand einer Person, die evaluative Bewertung umfasst kognitive, affektive und auch behaviorale Aspekte. Diese Definition hat mit der von Allport gemein, dass beide auf der Idee basieren, dass Einstellung einen inneren Zustand beschreibt, der zwischen einem Stimulus und einer darauf folgenden Reaktion interveniert (Eagly & Chaiken, 1993).

In der Literatur lassen sich noch viele weitere Definitionen ausmachen wie bspw. die von Pratkanis und Greenwald (1989), die Einstellung schlicht als „a person’s evaluation of an object of thought" (S. 247) definieren, oder diejenige von Zanna und Rempel (1988), die Einstellungen als „… items of knowledge in the form of evaluative summations“ (S. 330) verstehen. In nahezu allen Definitionsansätzen herrscht jedoch Einigkeit über den evaluativen bzw. bewertenden Charakter von Einstellungen (Haddock & Maio, 2019). Die Beziehungen der damit verknüpften affektiven, kognitiven und verhaltensbezogenen Aspekte werden in den folgenden Kapiteln weiter erläutert.

2.1.1.2. Inhalt und Entstehung von Einstellungen – Das Mehrkomponentenmodell der Einstellung

Zunächst gilt zu klären, wie Einstellungen in Reaktion auf ein Einstellungsobjekt zustande kommen. Ein prominentes Erklärungsmodell zur Entstehung von Einstellungen stellt das *Mehrkomponentenmodell* dar. Zurückzuführen u. a. auf die Arbeit von Zanna und Rempel (1988), trifft das Mehrkomponentenmodell die Annahme, dass sich die Entstehung von Einstellungen

auf drei Kategorien von Informationen und zugehörigen Verarbeitungsprozessen gründen kann. Konkret bestehen Einstellungen im Mehrkomponentenmodell aus kognitiven, affektiven und behavioralen Komponenten (siehe Abbildung 1). Einstellungen können dabei durch das gleichzeitige Ablaufen aller drei Kategorien von Verarbeitungsprozessen entstehen, jedoch auch nur auf zwei oder einer einzelnen Art von Prozessen beruhen (Zanna & Rempel, 1988).

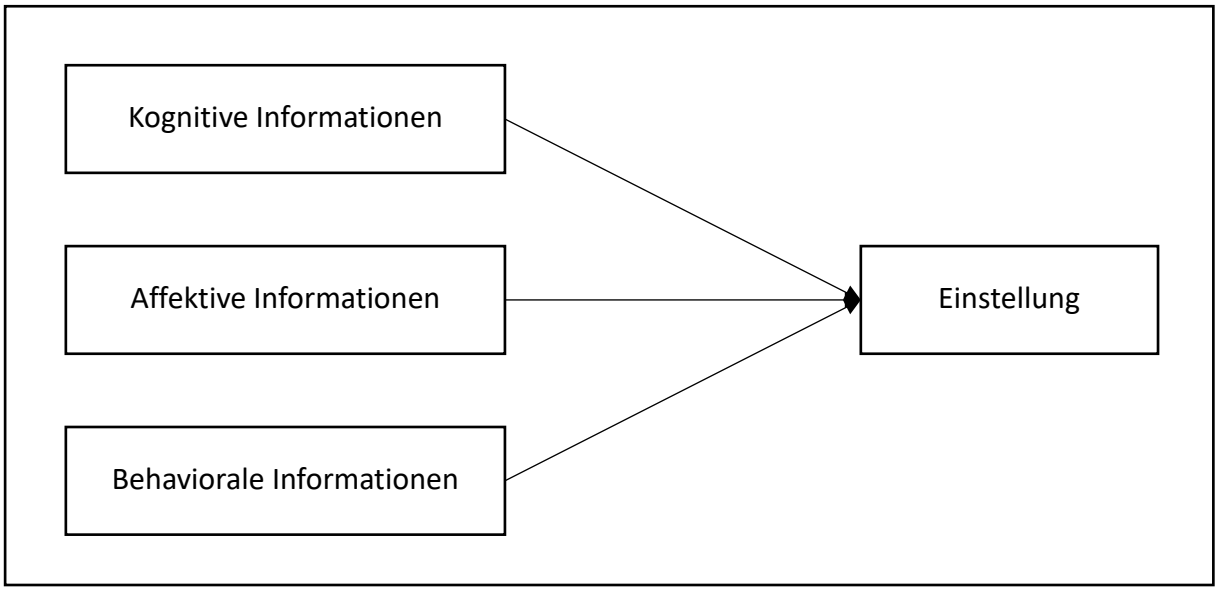


Abbildung 1: Mehrkomponentenmodell von Einstellung (Eagly & Chaiken, 1993; Maio & Haddock, 2010)

Die kognitive Komponente von Einstellungen bezieht sich auf Überzeugungen, Gedanken und Eigenschaften, die eine Person mit einem Einstellungsobjekt assoziiert (Maio & Haddock, 2010). Basierend auf direkten oder indirekten Erfahrungen mit einem Einstellungsobjekt bilden Menschen durch kognitive Verarbeitungs- bzw. Lernprozesse Überzeugungen bezüglich des betrachteten Objektes und entwickeln auf Basis einer Bewertung der Vorteilhaftigkeit des betrachteten Einstellungsobjektes eine Einstellung (Eagly & Chaiken, 1993). Einstellungen können alleine auf den mit einem Objekt assoziierten positiven oder negativen Eigenschaften und den darauf gründenden Überzeugungen basieren (Maio & Haddock, 2010). Ein aktuelles Beispiel aus dem Schulkontext wäre die Entscheidung einer Lehrkraft bei der Auswahl einer Online-Lernplattform durch die bewusste Gegenüberstellung der Eigenschaften verschiedener Plattformen, wie übersichtliche Gestaltung, Performance oder Verständlichkeit. Verschiedene Einstellungstheorien, die auch eine wichtige Grundlage des konzeptionellen Modells dieser Arbeit darstellen, basieren auf diesem kognitiven Ansatz. U. a. die Theory of Reasoned Action (TRA) sowie die Theory of Planned Behavior (TPB) ziehen kognitive Prozesse zur Erklärung der Entstehung von Einstellungen heran. Diese Theorien werden in Kapitel 2.1.2 näher

beleuchtet. Weitere in der Literatur prominente Modelle und Theorien, wie das *Information Processing Paradigm* (McGuire, 1968), das *Elaboration Likelihood Model* (Petty & Cacioppo, 1986) oder das *Heuristic-Systematic Model* (Chaiken, Libermann & Eagly, 1989) greifen diesen kognitiven Ansatz auf. Diese werden jedoch in dieser Arbeit nicht weiter behandelt, da der Fokus auf der Untersuchung des Verhältnisses von Einstellung und Verhalten liegt und diese Theorien im Gegensatz zu TRA und TPB dazu keinen direkten Erklärungsbeitrag leisten.

Unter der affektiven Einstellungskomponente werden die Gefühle und Emotionen in Zusammenhang mit dem Einstellungsobjekt verstanden. Bei einer Konfrontation mit dem Einstellungsobjekt beeinflussen Gefühle Einstellungen dabei durch affektive Reaktionen (Maio & Haddock, 2015). Im bereits oben genannten Beispiel wird eine Lehrkraft, die in einer positiven, entspannten Grundhaltung erstmals eine neue Lernplattform erproben kann, einem affektiven Verarbeitungsprozess folgend eher eine positive Einstellung gegenüber der Plattform entwickeln, als wenn sie dieselbe Lernplattform in einer negativen, z. B. von Stress geprägten Situation erprobt.

Die Entstehung von Einstellungen durch affektive Prozesse folgt u. a. den Ansätzen der *klassischen* und *operanten Konditionierung* nach Pavlov (1927) bzw. Skinner (1938) (Fischer, Asal & Krueger, 2014; Petty & Cacioppo, 1996). Eine der prominentesten empirischen Untersuchungen der Einstellungsbildung vor dem Hintergrund der klassischen Konditionierung stammt von Staats und Staats (1958), die dieser in zwei Experimenten auf den Grund gehen. Hierfür wurden Proband*innen im ersten Experiment verschiedene Ländernamen, im zweiten Männernamen, jeweils verknüpft mit unkonditionierten neutralen, positiven oder negativen verbalen Stimuli, präsentiert und im Anschluss daran u. a. gebeten, die Namen anhand eines semantischen Differenzials zu bewerten. Die Auswertung der Experimente verdeutlicht, dass die Bewertung in Abhängigkeit des präsentierten Stimulus entsprechend positiver oder negativer ausfällt. Generell wird im Kontext der Einstellungsbildung auch von *evaluativer Konditionierung* gesprochen (Haddock & Maio, 2014). Aus Sicht der evaluativen Konditionierung entsteht eine Einstellung als affektive Reaktion auf die gleichzeitige Darbietung eines positiven oder negativen Reizes (unkonditionierter Stimulus) zusammen mit dem betrachteten Einstellungsobjekt (konditionierter Stimulus) (Eagly & Chaiken, 1993; Maio & Haddock, 2010). Der Unterschied zur klassischen Konditionierung liegt hierbei darin, dass der unkonditionierte Stimulus die Entstehung einer inneren affektiven Reaktion im Sinne einer Einstellungsbildung hervorruft und nicht, wie bei der klassischen Konditionierung, die Manifestation eines bestimmten Verhaltens (Maio & Haddock, 2010). Je nach Wertigkeit bzw. Richtung des Reizes entsteht eine positive oder

negative Einstellung (Fischer et al., 2014). Im Rückgriff auf das bereits bekannte Beispiel wird eine Lehrkraft, die sich in ihrem gut ausgestatteten, gemütlichen Arbeitszimmer in Ruhe mit einer neuen Lernplattform auseinandersetzt, diese nach dem Modell der klassischen Konditionierung positiver bewerten, als wenn die Auseinandersetzung z. B. unterwegs in der Bahn unter dem Einfluss störender Umgebungsgeräusche und Gerüche erfolgt.

Dem Modell der operanten Konditionierung folgend bilden sich erwünschte Einstellungen durch Belohnung bei einstellungskonformem bzw. durch Bestrafung bei einstellungsdiskrepantem Verhalten (Fischer et al., 2014; Petty & Cacioppo, 1996). Die anfänglich skeptische Einstellung einer Lehrkraft gegenüber einer bestimmten Lernplattform könnte sich bspw. verbessern, wenn die Schüler*innen dem Unterricht der Lehrkraft, in dem die Plattform zum Einsatz kommt, ein positives Feedback geben.

Neben den Modellen der klassischen bzw. evaluativen und operanten Konditionierung stellt der *Mere-Exposure-Effekt*, zurückgehend auf die Arbeit von Zajonc (1968), einen weiteren Ansatz zur Erklärung der Entstehung von Einstellungen, basierend auf affektiven Reaktionen, dar. Der Mere-Exposure-Effekt beschreibt die Annahme, dass die reine wiederholte Darbietung eines Einstellungsobjektes ausreicht, eine positive Einstellung gegenüber dem Einstellungsobjekt zu entwickeln (Haddock & Maio, 2014; Maio & Haddock, 2010). Eine Reihe von Studien liefert empirische Hinweise darauf, dass die wiederholte Exposition einer Person gegenüber einem Einstellungsobjekt eine positive affektive Reaktion fördern kann (Haddock & Maio, 2014), wobei zu beachten gilt, dass die reine Darbietung eines Einstellungsobjektes i. d. R. nicht als alleinige Ursache affektiver Reaktionen gesehen werden kann (Fishbein & Ajzen, 1975).

Die behaviorale oder auch konative Komponente von Einstellungen bezieht sich auf frühere Verhaltensweisen im Zusammenhang mit einem Einstellungsobjekt (Haddock & Maio, 2014). Behaviorale Prozesse der Einstellungsbildung spiegeln sich dabei in der These, dass Einstellungen durch vergangenes Verhalten oder Erfahrungen beeinflusst werden (Maio & Haddock, 2015). Zurückzuführen ist diese Annahme auf die *Selbstwahrnehmungstheorie* (*Self-Perception Theory*) von Bem (1972). Gemäß der Selbstwahrnehmungstheorie betrachten und analysieren Menschen ihr eigenes Verhalten ebenso, wie sie das Verhalten anderer Menschen analysieren. Aus diesen Beobachtungen des eigenen Verhaltens oder den Umständen, in denen ein Verhalten gezeigt wurde, leiten sie ihre Einstellungen, Gefühle und andere innere Zustände ab (Bem, 1972). Insbesondere bei schwachen oder schwer zugänglichen bzw. unbewusst vorliegenden Einstellungen neigen Menschen demnach dazu, ihr eigenes vergangenes Verhalten zu

analysieren. Sie bilden ihre Einstellung rückwirkend, indem sie Informationen aus dieser Verhaltensanalyse interpretieren. Die Verhaltensinterpretation wird dabei von der Frage geleitet, welche Einstellung dem gezeigten Verhalten zugrunde lag (Bem, 1972; Fischer et al., 2014). In dem hier verwendeten schulischen Beispiel würde demzufolge eine Lehrkraft, die in ihrem Unterricht häufig eine digitale Lernplattform nutzt, weil es an ihrer Schule so üblich ist, wahrscheinlich in einer Befragung eine positive Einstellung zu digitalen Lernplattformen angeben, auch wenn sie zuvor noch nicht explizit darüber nachgedacht hat.

Da das Verhalten als Triebkraft der Einstellungsbildung im Fokus der Selbstwahrnehmungstheorie steht, wird diese in der Literatur im Kontext der Einstellungsbildung zwar als behavioristischer Ansatz kategorisiert, dennoch führen durch die Verarbeitung und Interpretation des Verhaltens auch kognitive Prozesse zur Bildung einer Einstellung (Eagly & Chaiken, 1993; Fischer et al., 2014). Vor dem Hintergrund des Zusammenspiels kognitiver und behavioraler Prozesse kann auch Festingers (1957, 1964) *Theorie der kognitiven Dissonanz* zur Erklärung der Entstehung verhaltensbasierter Einstellungen herangezogen werden, eine Theorie, in der Kognitionen eine zentrale Rolle spielen (Maio & Haddock, 2010). Der Dissonanztheorie folgend führt ein Verhalten bzw. ein verhaltensbezogenes kognitives Element, das von der eigenen Einstellung abweicht, zu diskrepanten Kognitionen. Menschen, die nach kognitiver Konsistenz streben, versuchen die dadurch entstehende aversive Spannung zu reduzieren, indem sie ihre Einstellung ihrem Verhalten anpassen (Festinger, 1957). Bspw. sieht sich eine Lehrkraft im Zuge der Homeschooling Verpflichtung während der Corona Pandemie zur Aufrechterhaltung des Unterrichts zum Einsatz einer bestimmten Online-Lernplattform genötigt, auch wenn sie Vorbehalte gegen derartige digitale Systeme hegt. Trotz der bestehenden Vorbehalte wird die erzwungene Nutzung dazu führen, dass sich die Einstellung der Lehrkraft zum Positiven verändert. Weitere Strategien zur Dissonanzreduktion bestehen bspw. in der Anpassung des Verhaltens oder der individuellen Bedeutsamkeit bestimmter Kognitionen (Festinger, 1957; Maio & Haddock, 2010). Diese sind jedoch im speziellen Kontext der Einstellungsbildung bzw. -anpassung nicht weiter relevant und werden daher an dieser Stelle nicht näher behandelt.

Die Auseinandersetzung mit dem eigenen Verhalten durch kognitive Prozesse in der Theorie der kognitiven Dissonanz verdeutlicht, dass es durchaus nicht trivial ist, die verschiedenen Prozesse zur Herausbildung oder Anpassung von Einstellungen klar zu trennen. Dies unterstreicht auch die von Maio und Haddock (2010) im Hinblick auf das Mehrkomponentenmodell aufgeworfenen Frage, ob sich die einzelnen Komponenten tatsächlich unterscheiden. Die Autoren kommen basierend auf den Erkenntnissen der Arbeit von Breckler (1984) zu dem Schluss, dass

es durchaus wissenschaftliche Evidenz dafür gibt, dass sich affektive, kognitive und behaviorale Komponenten von Einstellungen voneinander abgrenzen lassen. Dies bedeutet jedoch nicht, dass die verschiedenen Prozesse vollkommen unabhängig voneinander sind. Häufig spiegeln sich positive Überzeugungen gegenüber einem Einstellungsobjekt auch in positiven Gefühlen und Verhaltensweisen gegenüber dem Einstellungsobjekt wider (Haddock & Maio, 2014). Jedoch gibt es auch Situationen, in denen sich die evaluativen Implikationen unterscheiden (Maio & Haddock, 2015). Als Beispiel für mögliche widersprüchliche evaluative Reaktionen gegenüber einem Einstellungsobjekt betrachten wir eine Lehrkraft, die den Einsatz einer Online-Lernplattform als wünschenswert und förderlich für ihren Unterricht ansieht (positive Kognition), aber ängstlich und unsicher im Umgang mit Technologien ist (negativer Affekt) und auch bereits schlechte Erfahrungen bei der Nutzung digitaler Instrumente im Unterricht gemacht hat (erfahrungsbasierte negative Konation).

Generell müssen Einstellungen nicht zwangsläufig alle drei Aspekte umfassen, sondern können auch primär oder ausschließlich auf einem der drei möglichen Prozesse basieren. Häufig bilden sich Einstellungen jedoch durch eine Mischung verschiedener Prozesse (Eagly & Chaiken, 1993).

Zusammenfassend lässt sich hinsichtlich der Entstehung von Einstellungen festhalten, dass diese sowohl auf affektiven, kognitiven als auch behavioralen Prozessen basieren können, die sich auf der einen Seite zwar empirisch unterscheiden lassen, auf der anderen Seite dennoch nicht komplett unabhängig voneinander sind, und somit häufig verschiedene Prozesse zum Herausbilden einer Einstellung beitragen. Manche Einstellungen basieren dabei eher auf Kognitionen, andere eher auf affektiven Prozessen (Haddock & Maio, 2014). So richten die beschriebenen Theorien ihre Aufmerksamkeit i. d. R. nicht ausschließlich auf eine Komponente von Einstellungen, sondern setzen meist einen speziellen Fokus auf den besonderen Einfluss bestimmter Prozesse, während andere Prozesse eine untergeordnete Rolle spielen. Generell existieren viele weitere Modelle und Theorien zur Entstehung und Veränderung von Einstellungen, deren Ausführung jedoch keinen relevanten Beitrag zum Erkenntnisinteresse dieser Arbeit leisten würde und daher nicht weiter thematisiert werden. Ein guter Überblick hierzu findet sich bspw. bei Eagly und Chaiken (1993).

2.1.1.3. Drei-Komponenten-Modell nach Rosenberg und Hovland

Wie im zweiten Abschnitt dieses Kapitels im Rahmen des Mehrkomponentenmodells dargestellt, entstehen Einstellungen durch die Wahrnehmung und Verarbeitung kognitiver, affektiver oder behavioraler Stimuli, die dem Einstellungsobjekt zuzuordnen sind. Erst durch die evaluative Reaktion auf diese Stimuli, die sich in Form einer zustimmenden oder ablehnenden Reaktion ausdrückt, wird eine Einstellung beobachtbar bzw. messbar (Eagly & Chaiken, 1993). Einstellungen sind dabei Prädispositionen, in bestimmter Weise auf eine bestimmte Art von Stimulus zu reagieren (Rosenberg & Hovland, 1969). Die Dimensionalität von Einstellungen ist eine in der Literatur kontrovers diskutierte Frage (Dillon & Kumar, 1985). Frühe Arbeiten zur Erforschung von Einstellungen sahen diese meist als eindimensionales Konstrukt, wie bspw. Thurstone (1929), dessen Skala zur Einstellungsmessung primär auf die Erfassung von Gefühlen und Emotionen der Befragten zu einem Einstellungsobjekt abzielt, und welcher Einstellung als affektives Konstrukt begreift (Rosenberg & Hovland, 1969). Andere Arbeiten, wie die von Bagozzi und Burnkrant (1979) oder Zajonc und Markus (1982), unterstellen Einstellungen eine zweidimensionale Struktur, basierend auf den Konstrukten Kognition und Affekt (Zanna & Rempel, 1988). Die Arbeit von Rosenberg und Hovland (1969) grenzt sich von einer ein- bzw. zweidimensionalen Sichtweise ab, indem die Autoren eine komplexere Struktur von Einstellungen annehmen. Ebenso wie bei dem einen Einstellungsobjekt zugehörigen Stimulus unterscheidet das *Drei-Komponenten-Modell* nach Rosenberg und Hovland (1969) auch auf Seiten der evaluativen Reaktion drei Kategorien (*response categories*). Das Modell differenziert zwischen kognitiven, affektiven und behavioralen Reaktionen, die sich jeweils sowohl verbal als auch nonverbal äußern können (siehe auch Abbildung 2) (Eagly & Chaiken, 1993; 1998).

Die erste Kategorie einstellungsbezogener Reaktionen – der Affekt – umfasst Gefühle, Emotionen und Stimmungen in Bezug auf ein Einstellungsobjekt, ebenso wie durch dieses hervorgerufene Reaktionen des sympathischen Nervensystems (Eagly & Chaiken, 1993). Affekte sind hierbei kurzfristige angenehme oder unangenehme emotionale oder gefühlsmäßige Zustände, die durch mehr oder weniger positive bzw. negative Bewertungen des Einstellungsobjektes ausgelöst werden (Eagly & Chaiken, 2005). Die affektive Reaktion bewegt sich somit auf einer evaluativen Dimension in einem Bereich von extrem negativ bis extrem positiv (Eagly & Chaiken, 1998). Darüber hinaus kann zwischen Affekten unterschieden werden, die direkt durch den Umgang mit einem Einstellungsobjekt ausgelöst werden, und solchen, die eher vage mit dem Einstellungsobjekt assoziiert werden (Eagly & Chaiken, 2005). Messbar werden affektive Reaktionen zum einen durch den verbalisierten Ausdruck von Emotionen und Gefühlen im

Sinne von Aussagen bezüglich Zuneigung oder Ablehnung im Hinblick auf ein Einstellungsobjekt. Darüber hinaus spiegeln sich affektive Reaktionen in direkt beobachtbaren bzw. messbaren physiologischen Reaktionen, wie Herzschlag oder Blutdruck, wider (Eagly & Chaiken, 1998; Rosenberg & Hovland, 1969). Bewertet eine Person ein Einstellungsobjekt positiv, zeigt sie generell wahrscheinlich auch eine positive affektive Reaktion auf das Einstellungsobjekt bei einer gleichzeitig geringen Wahrscheinlichkeit einer negativen Reaktion und umgekehrt (Eagly & Chaiken, 1993). Diesbezüglich wäre eine Lehrkraft, die im Hinblick auf die Nutzung einer Online-Lernplattform eher verunsichert oder sogar ängstlich reagiert, ein Beispiel für eine negative affektive Einstellungsreaktion.

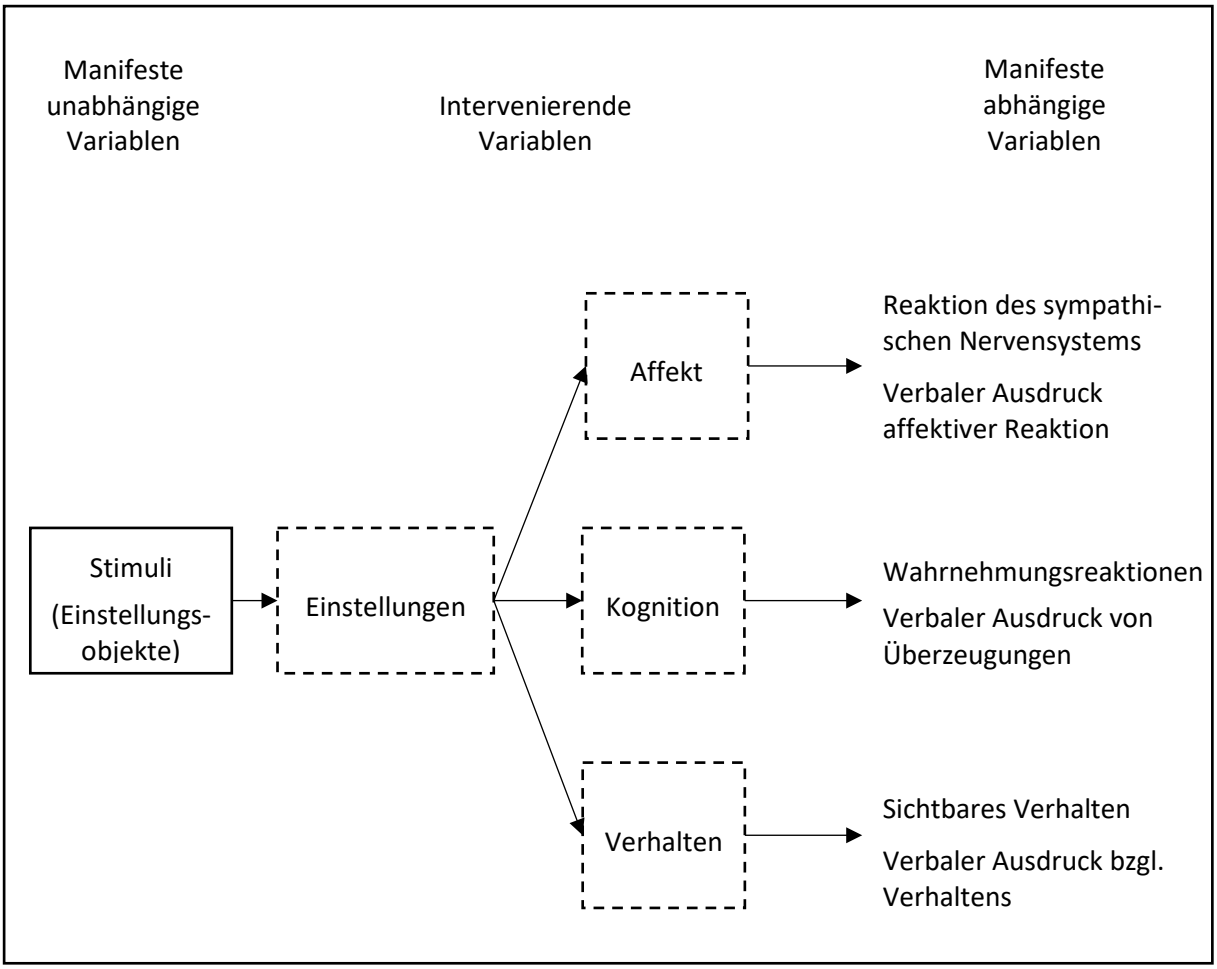


Abbildung 2: Drei-Komponenten-Modell von Einstellung (Darstellung nach Rosenberg & Hovland, 1969)

Die zweite Kategorie – die Kognition – bezieht sich auf die Gedanken und Vorstellungen einer Person in der Wahrnehmung des betrachteten Einstellungsobjektes. Diese kognitive Bewertungskategorie wird dabei häufig als *Überzeugung (belief)* konzeptualisiert (Eagly & Chaiken, 1993, 1998). Überzeugungen repräsentieren die Informationen, die eine Person über ein

(Einstellungs-)Objekt besitzt. Eine Überzeugung setzt demnach ein Objekt mit bestimmten Eigenschaften in Verbindung, die eine Person dem Objekt zuschreibt (Fishbein & Ajzen, 1975). Diese mit dem Einstellungsobjekt assoziierten Attribute umfassen die Bewertungen des Einstellungsobjektes, die, ebenso wie affektive Reaktionen, auf einem Kontinuum von negativ bis positiv abgebildet werden können (Eagly & Chaiken, 1993). Eine Lehrkraft mit der Überzeugung, der Einsatz einer Online-Lernplattform (Einstellungsobjekt) im Unterricht sei kompliziert (Attribut), kann dabei als Beispiel einer negativen Attributzuschreibung zu einem Objekt herangezogen werden. Generell können einige Überzeugungen dabei sehr spezifisch, andere dagegen eher abstrakt sein (Eagly & Chaiken, 1998). Kognitive Reaktionen manifestieren sich letztlich in verbal oder nonverbal zum Ausdruck gebrachten messbaren Indikatoren. Verbale Indikatoren zeigen sich bspw. in schriftlich oder mündlich geäußerten Überzeugungen hinsichtlich der Attribute eines Einstellungsobjektes, nonverbale Indikatoren beziehen sich auf wahrnehmungsbezogene Reaktionen (Rosenberg & Hovland, 1969).

Die dritte Kategorie einstellungsbezogener Reaktionen – die Konation – betrifft das Verhalten einer Person im Hinblick auf ein Einstellungsobjekt (Eagly & Chaiken, 1993). Neben dem tatsächlichen Verhalten gegenüber einem Einstellungsobjekt umfasst die konative Komponente auch Handlungsintentionen, die wiederum nicht immer notwendigerweise in einem Verhalten resultieren (Eagly & Chaiken, 1998). Messbare Indikatoren konativer Einstellungsreaktionen stellen zum einen das direkt beobachtbare Verhalten einer Person gegenüber einem Einstellungsobjekt dar und lassen sich zum anderen aus der verbal geäußerten Verhaltensabsicht oder aus Berichten hinsichtlich vergangenen Verhaltens ableiten (Rosenberg & Hovland, 1969). Die Ausprägungen konativer Einstellungsreaktionen können dabei, ebenso wie die der anderen Komponenten, von extrem negativ bis extrem positiv reichen. Generell führt die positive Bewertung eines Einstellungsobjekts zu förderlichen Verhaltensweisen bezüglich des Objekts (Eagly & Chaiken, 1993). Eine Lehrkraft bspw., die eine bestimmte Online-Lernplattform gut findet, berichtet ihren Kolleginnen und Kollegen darüber und wirbt für das Instrument.

Zusammenfassend kann festgehalten werden, dass das Drei-Komponenten-Modell durch seine Abgrenzung zur eindimensionalen Struktur von Einstellungen eine Weiterentwicklung der Einstellungstheorien darstellt. Zwar ist das Modell nicht ausreichend, um die gesamte Komplexität von Einstellungen allgemein und im Speziellen im Kontext der Akzeptanz von VERA adäquat zu erfassen, dennoch dient das Modell als wichtige heuristische Grundlage zum Verständnis einstellungsbezogener Reaktionen, da auch die in den folgenden Kapiteln vorgestellten Modelle und Theorien auf dieser Dreigliedrigkeit aufbauen. Das Modell betrachtet alle drei

Komponenten als gleichwertige bzw. auf einer Ebene angeordnete Facetten von Einstellung, die sich auf jede dieser drei Weisen widerspiegeln kann. Konkret werden Kognition, Affekt und Konation als Faktoren erster Ordnung und die Gesamtbewertung (*overall evaluation*) bzw. die Einstellung als übergeordneter Faktor zweiter Ordnung interpretiert (Ajzen, 1989). Die in den folgenden Kapiteln vorgestellten Theorien lösen sich von dieser Sichtweise und betrachten die behaviorale Komponente, welche im Drei-Komponenten-Modell inhärenter Bestandteil einer Einstellung ist, als ein der Einstellung nachgelagertes Konstrukt. Ziel dieser Verhaltenstheorien ist, das Verhalten von Individuen durch Einstellungen und andere Faktoren zu erklären.

2.1.2. Die Beziehung von Einstellung und Verhalten – Verhaltenstheorien

2.1.2.1. Einstellung und Verhalten

Da es eines der Ziele dieser Arbeit ist, die Nutzung bzw. die Nicht-Nutzung der VERA-Rückmeldungen durch Lehrkräfte zu erklären, ist es von Bedeutung, das Verhältnis von Einstellungen und Verhalten zu verstehen. Für diesen Zweck stellen *Verhaltenstheorien* eine wichtige theoretische Grundlage dar, da diese speziell darauf abzielen, Verhalten zu erklären. Verhalten wird dabei nicht als Teil einer Einstellung, sondern als Konsequenz von Einstellung betrachtet. Ausgangspunkt der Erforschung der Einstellungs-Verhaltens-Relation bilden *Konsistenzmodelle* und die Annahme einer konsistenten und widerspruchsfreien Beziehung zwischen Einstellung und Verhalten. Die Grundlage dieser Konsistenzannahme bildet die Auffassung, dass Individuen nach kognitiver Konsistenz und Ausgewogenheit streben. Demzufolge ergibt sich aus einer positiven Einstellung eine positive (Verhaltens-)Reaktion, eine negative Einstellung führt zu einer negativen (Verhaltens-)Reaktion (Pratkanis & Greenwald, 1989).

Zu den bekanntesten Konsistenztheorien zählt, neben der bereits in Kapitel 2.1.1 erwähnten Dissonanztheorie (Festinger, 1957), u. a. die *Balance Theorie* (Heider, 1946). Ähnlich wie in der Dissonanztheorie, streben Individuen gemäß der Balance Theorie nach einem ausbalancierten kognitiven System und spannungsfreien Beziehungen zu anderen Personen und Objekten. Im Hinblick auf die Beziehung von Einstellung und Verhalten bedeutet dies, dass Verhaltensweisen, die der eigenen Einstellung widersprechen, als kognitiv inkonsistent empfunden werden, wodurch eine innere Spannung erzeugt wird, die als unangenehm empfunden wird. Das Streben nach Konsistenz bzw. Balance aller eine bestimmte Einheit – bspw. ein

Einstellungsobjekt – betreffenden Kognitionen führt dabei zu einem einstellungskonformen Verhalten (Eagly & Chaiken, 1993; Petty & Cacioppo, 1996).

Herrschte in den Veröffentlichungen zu den verschiedenen Konsistenztheorien noch weitestgehender Konsens bezüglich einer konsistenten 1:1 Beziehung von Einstellung und Verhalten, ergaben sich in den 60er Jahren neue Impulse zur Erforschung der Einstellungs-Verhaltens-Relation (Fazio, 1986). U. a. die Arbeiten von Wicker (1969) und Festinger (1964) stellten die bis dahin vorherrschende Annahme einer perfekten Verhaltensvorhersage durch Einstellungen in Frage und unterstellten eine komplexere Beziehung. Angestoßen durch die Arbeit von Wicker (1969) wurde die Beziehung von Einstellung und Verhalten vielfach weiter untersucht, mit der Erkenntnis, dass Einstellungen durchaus unter bestimmten Bedingungen Verhalten voraussagen, die Annahme einer 1:1 Beziehung jedoch zu kurz gegriffen ist (Haddock & Maio, 2014). Gründe, warum Verhaltensweisen einer Person nicht zwangsläufig ihren Einstellungen entsprechen, liegen nicht nur darin, dass häufig auch andere konkurrierende Einstellungen existieren, die ihrerseits das Verhalten beeinflussen, sondern auch im gleichzeitigen Vorhandensein weiterer Einflussfaktoren, wie Wahrnehmungskonstrukte oder sonstige externe Einflüsse und Restriktionen (Bohner & Schwarz, 2001; Felser, 2015). Bspw. kann die Nicht-Nutzung einer Online-Lernplattform trotz positiver Einstellung einer Lehrkraft auf eine mangelnde technische Ausstattung der Schule zurückzuführen sein. Die in den folgenden Kapiteln vorgestellten Verhaltenstheorien beschäftigen sich mit eben diesen diversen, neben Einstellungen existierenden Antezedenzien von Verhalten und der Erklärung von einstellungsbasierten Verhaltensweisen.

2.1.2.2. Theory of Reasoned Action (TRA)

Die erste für diese Arbeit relevante Verhaltenstheorie ist Fishbein und Ajzens (1975) Theory of Reasoned Action (TRA). Die Autoren bemängeln an dem Drei-Komponenten-Modell von Rosenberg und Hovland (1969), dass dieses die Relation von Einstellung und Verhalten nicht in ausreichendem Maße erklären kann. Darüber hinaus beschreiben sie die Problematik vieler Studien, die gemessenen Konstrukte korrekt zu identifizieren bzw. zu definieren. Häufig würden Konstrukte als Einstellungen betrachtet, die eigentlich keine Einstellungen repräsentieren (z. B. Wichtigkeit) und umgekehrt Konstrukte, die tatsächlich Einstellungen darstellen, nicht als Einstellungen definiert (z. B. wahrgenommene Konsequenzen). Auch im Hinblick auf das Konstrukt Verhalten sehen die Autoren oftmals eine gewisse Uneindeutigkeit. Demnach betrachten viele Studien die Beziehung von Einstellung und Verhalten, messen jedoch nur Verhaltensintentionen und nicht das tatsächliche Verhalten. Getrieben von der Kritik an bis dahin

vorherrschenden Modellen und insbesondere dem Drei-Komponenten-Modell nimmt die TRA eine kausale Neustrukturierung der drei Einstellungskomponenten vor. Kognitive Elemente beeinflussen dabei affektive bzw. evaluative Elemente, welche wiederum das Verhalten beeinflussen. Die Theorie liefert einen, u. a. auf Einstellungen basierenden, Ansatz zur Erklärung begründeten und überlegten Verhaltens und betrachtet Einstellungen dabei als einen einer Handlung vorausgehenden Status. Grundlegende Prämisse der Theorie ist dabei die Annahme, dass Menschen sich rational verhalten und ihre Entscheidungen und ihr Verhalten auf systematische Informationsverarbeitung gründen (Ajzen & Fishbein, 1980). Der konkrete Aufbau des Modells ist aus Abbildung 3 ersichtlich und wird im Folgenden beschrieben. An dieser Stelle ist noch anzumerken, dass im Kontext der TRA und generell in dieser Arbeit die Begriffe Verhalten und Handlung synonym verwendet werden.

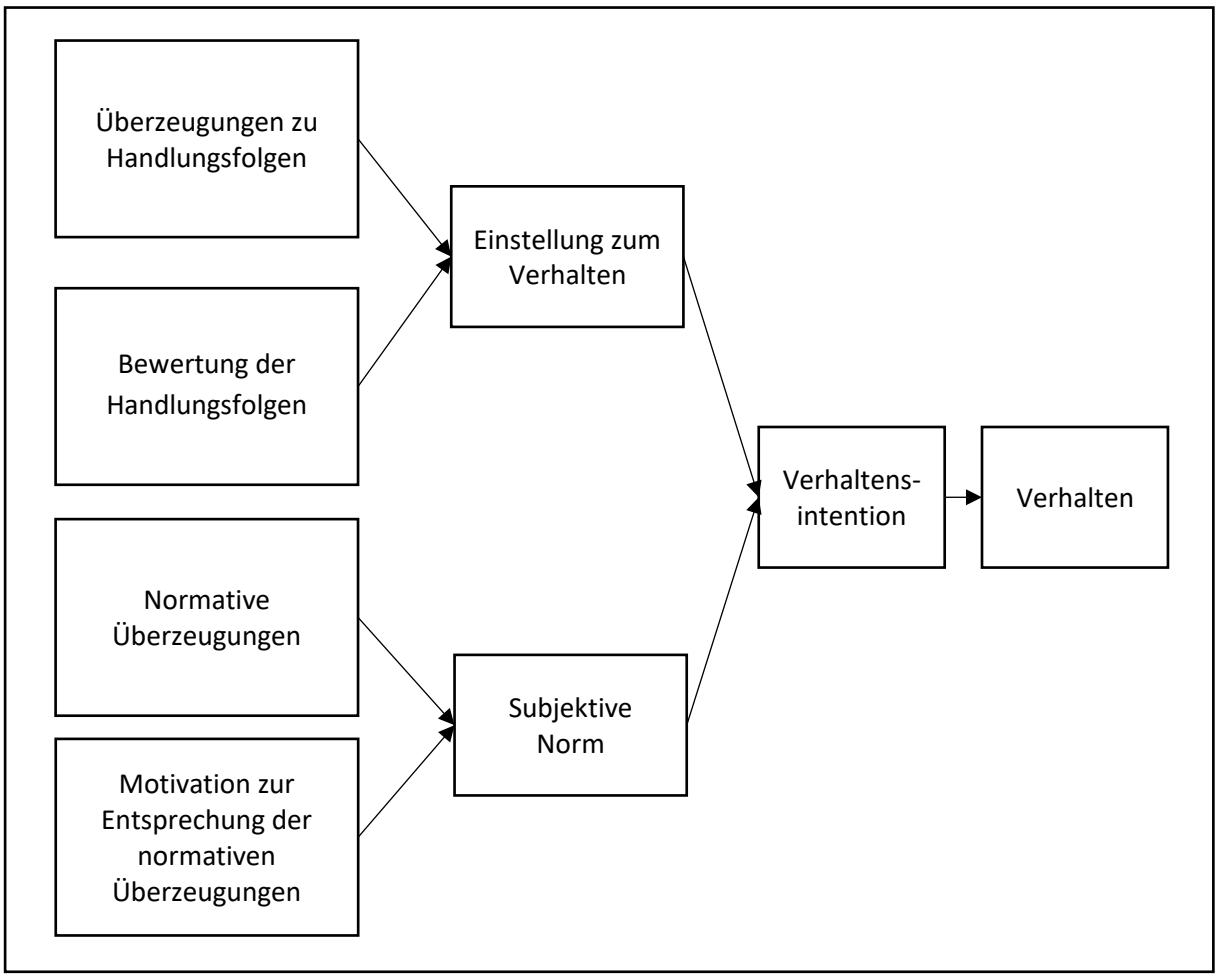


Abbildung 3: Theory of Reasoned Action (Fishbein & Ajzen, 1975); Abbildung in Anlehnung an (F. D. Davis et al., 1989; Eagly & Chaiken, 1993)

Bei der Betrachtung der TRA fällt zunächst auf, dass das Modell die Verhaltenskomponente differenziert betrachtet und zwischen tatsächlichem Verhalten und Verhaltensintention

unterscheidet. Die Verhaltensintention, genauer die subjektive Wahrscheinlichkeit, ein bestimmtes Verhalten in Bezug auf das Einstellungsobjekt zu zeigen, ist dem eigentlichen Verhalten dabei unmittelbar vorgelagert. Das Modell beinhaltet die Annahme, dass das Vorhaben, ein bestimmtes Verhalten zu zeigen oder eben nicht zu zeigen, einen bzw. den einzigen unmittelbaren Einflussfaktor des tatsächlichen Handelns darstellt (Ajzen & Fishbein, 1980). Einstellung wirkt dadurch nicht direkt auf das tatsächliche und beobachtbare Verhalten, sondern lediglich über eine Verhaltensabsicht als intervenierende Variable. Durch die Unterscheidung zwischen Verhaltensintention und tatsächlichem Verhalten in zwei behaviorale Konstrukte grenzt sich die TRA von früheren Modellen ab. In einem Beispiel aus der Schulpraxis würde dies bedeuten, dass eine Lehrkraft, die plant, im kommenden Schuljahr in der achten Klasse eine Online-Lernplattform einzusetzen, dieses Vorhaben mit einer relativ hohen Wahrscheinlichkeit umsetzen wird. Verhaltensintention wird dabei nicht, wie häufig in älteren Modellen, einfach unter Einstellung subsumiert bzw. als behaviorale Komponente von Einstellung angesehen. In der TRA ist die Intention kein Teil der Einstellung selbst, sondern dieser nachgelagert (Fishbein & Ajzen, 1975).

Der Theorie zufolge bestimmen nun zum einen Einstellungen *(attitudes)* und zum anderen subjektive Normen *(subjective norms)* die Intention, eine Handlung durchzuführen oder zu unterlassen (Fishbein & Ajzen, 1975). Einstellungen repräsentieren dabei einen individuellen Bestimmungsfaktor; subjektive Normen spiegeln soziale Einflüsse wider (Ajzen & Fishbein, 1980). Fishbein und Ajzen (1975) definieren Einstellung als individuelles positives oder negatives Gefühl einer Person gegenüber einem Einstellungsobjekt. Einstellungen sind hierbei auf einer bipolaren Dimension angeordnet und drücken eine evaluative bzw. affektive Bewertung aus. Einstellungen repräsentieren in der TRA jedoch keine Einstellungen zu (Einstellungs-)Objekten (z. B. die Einstellung einer Lehrkraft zu Online-Lernplattformen), sondern Einstellungen zum Verhalten *(attitude toward behavior)* im Hinblick auf das (Einstellungs-)Objekt. Dies bezieht sich konkret auf die Frage, ob eine Person ein bestimmtes Verhalten als gut oder schlecht bzw. positiv oder negativ bewertet (Maio & Haddock, 2015). In dem genannten Beispiel wäre die Einstellung zum Verhalten repräsentiert durch die Einstellung einer Lehrkraft zur Nutzung einer Online-Lernplattform für den eigenen Unterricht. Eine Lehrkraft mit einer positiven Einstellung zur Nutzung einer Online-Lernplattform würde demnach eher planen, diese im nächsten Schuljahr im Unterricht einzusetzen, als eine Lehrkraft mit einer negativen Einstellung. Die Einstellung zu einem (Einstellungs-)Objekt und die Einstellung gegenüber der Nutzung bzw. zu einem Verhalten gegenüber einem (Einstellungs-) Objekt müssen sich dabei nicht

zwangsläufig entsprechen. Eine positive Einstellung zur Nutzung eines Einstellungsobjektes kann durchaus zu einer positiven Handlungsintention und einem entsprechenden Verhalten führen, obwohl die Einstellung dem Einstellungsobjekt selbst gegenüber eher negativ ist, und umgekehrt (Ajzen & Fishbein, 1980). Um auf das bereits bekannte Beispiel zurückzugreifen, könnte eine Lehrkraft mit einer negativen Einstellung zu Online-Lernplattformen durchaus eine positive Einstellung zur Nutzung dieser Instrumente besitzen, wenn sie bspw. annimmt, durch den Einsatz dieser modernen Technik ihre Reputation im Kollegium, bei Vorgesetzten oder bei ihren Schüler*innen steigern zu können.

Subjektive Norm bezeichnet die individuelle Auffassung dessen, wie andere, insbesondere nahestehende Personen, ein bestimmtes Verhalten in Bezug auf das Einstellungsobjekt bewerten und berücksichtigt dabei den Einfluss des individuellen sozialen Umfeldes auf das Verhalten einer Person (Fishbein & Ajzen, 1975). Die subjektive Norm spiegelt somit den wahrgenommenen sozialen Druck wider, ein bestimmtes Verhalten auszuüben oder zu unterlassen (Maio & Haddock, 2015). In dem hier gewählten Beispiel könnte die Erwartung der Schulleitung, dass die Lehrkräfte digitale Unterrichtstools einsetzen, die subjektive Norm repräsentieren und dazu führen, dass eine Lehrkraft den Einsatz einer Online-Lernplattform für den zukünftigen Unterricht in Betracht zieht.

Die beiden Einflussfaktoren der Verhaltensintention, Einstellungen und subjektive Normen, werden wiederum u. a. jeweils durch bestimmte Überzeugungen (beliefs) beeinflusst. Allgemein gesprochen repräsentieren Überzeugungen die Verknüpfung eines Objektes, auf welches sich die Überzeugung bezieht, mit bestimmten Merkmalen. Die Begriffe Objekt bzw. Merkmal stehen dabei nicht zwangsläufig für Objekte im Sinne eines Gegenstandes und dessen direkt wahrnehmbare Eigenschaften, sondern auch für alle weiteren denkbaren Aspekte wie Werte, Konzepte oder sonstige Charakteristika. Eine Person nimmt bspw. an, bestimmte Eigenschaften zu besitzen (z. B. hält sich für lustig), vermutet bestimmte Konsequenzen für ein bestimmtes Verhalten (z. B. lernen führt zu guten Noten), schreibt Objekten bestimmte Eigenschaften zu (z. B. eine Banane ist krumm), etc. Konkret bemisst sich eine Überzeugung durch die subjektive Wahrscheinlichkeit, dass eine Verknüpfung zwischen einem Objekt und einem weiteren Objekt, bzw. Merkmal besteht. Die Stärke einer Überzeugung bemisst sich dabei daran, wie hoch diese Wahrscheinlichkeit eingeschätzt wird (Fishbein & Ajzen, 1975). Überzeugungen bilden kognitive Prozesse ab, die in der TRA der Einstellungsbildung vorausgehen. Im Gegensatz zum Drei-Komponenten-Modell, in welchem Kognitionen eine Facette von Einstellungen darstellen, sind Kognitionen in der TRA somit kausales Antezedens von Einstellung (Ajzen, 1989).

In der TRA sind zwei Kategorien von Überzeugungen von Bedeutung: zum einen *normative Überzeugungen (normative beliefs)*, die die Herausbildung subjektiver Normen beeinflussen, zum anderen *verhaltensbezogene Überzeugungen (behavioral beliefs)*, welche die Voraussetzung zur Entstehung von Einstellungen darstellen (Fishbein & Ajzen, 1975). Einstellungen werden bestimmt von der Erwartung einer Person, dass ein Verhalten zu einem bestimmten Ergebnis führt, und dem Wert, der diesem Ergebnis zugeschrieben wird (Fishbein & Ajzen, 1975; Haddock & Maio, 2014). Diese Annahme folgt der Logik von *Erwartung-mal-Wert-Modellen* (*Expectancy-Value Model*), die im folgenden Abschnitt kurz erläutert werden (Fishbein & Ajzen, 1975).

Erwartung-mal-Wert-Modelle definieren Einstellungen als eine Funktion der Überzeugungen zu einem Einstellungsobjekt. Überzeugungen ergeben sich dabei aus der Summe der erwarteten Werte (*expected values*) der einzelnen Attribute, die dem Einstellungsobjekt zugeschrieben werden. Die erwarteten Werte wiederum haben jeweils eine Erwartungs- und eine Wertkomponente. Die Erwartungskomponente beschreibt die subjektive Wahrscheinlichkeit, dass ein Einstellungsobjekt ein Attribut besitzt; die Wertkomponente bezieht sich auf die Bewertung des jeweiligen Merkmals. Eine Einstellung ergibt sich aus der Produktsumme der jeweiligen Erwartungs- und Wertkomponenten der einzelnen Attribute (Eagly & Chaiken, 1993). Bspw. nimmt eine Lehrkraft eine neue Online-Lernplattform einerseits als sehr innovativ und neuartig wahr, die Bedienung jedoch als nicht einfach verständlich. Die Einstellung ergibt sich nun aus der jeweiligen subjektiven Wahrscheinlichkeit, dass das Einstellungsobjekt diese Attribute tatsächlich besitzt – hohe Wahrscheinlichkeit der Innovativität und geringe Wahrscheinlichkeit der Verständlichkeit – und der in diesem Fall positiven Bewertung der beiden Eigenschaften.

Gemäß der TRA ergibt sich die Einstellung nun konkret aus der Produktsumme verhaltensbezogener Überzeugungen (behavioral beliefs), also den Überzeugungen hinsichtlich der Konsequenzen eines bestimmten Verhaltens und der Bewertung der entsprechenden Handlungsfolgen (*evaluation*) (Fishbein & Ajzen, 1975). Die Erwartung, dass der Einsatz einer Online-Lernplattform einen abwechslungsreichen Unterricht fördert (behavioral belief), würde zusammen mit der Bewertung, dass ein abwechslungsreicher Unterricht erstrebenswert ist (*evaluation*), zu einer positiven Einstellung der im Beispiel betrachteten Lehrkraft hinsichtlich des Einsatzes einer Online-Lernplattform führen. Im Hinblick auf externe Stimuli, wie Eigenschaften des betrachteten Einstellungsobjektes, Umwelteinflüsse, demografische Variablen oder individuelle Charakteristika, bedeutet der Erwartung-mal-Wert-Ansatz der TRA, dass diese externen Faktoren

Einstellungen nicht direkt, sondern nur indirekt durch Veränderungen in den Überzeugungsstrukturen einer Person beeinflussen können (F. D. Davis et al., 1989; Fishbein & Ajzen, 1975).

Die subjektive Norm ergibt sich wiederum aus der Produktsumme normativer Überzeugungen (normative beliefs) und der Motivation diesen zu entsprechen (*motivation to comply*). Normative Überzeugungen betreffen die Überzeugungen einer Person hinsichtlich der Erwartungen einer bestimmten Referenzgruppe oder eines Individuums, wonach die Person ein bestimmtes Verhalten zeigen oder nicht zeigen soll. Der Faktor Motivation bezieht sich auf die Motivation, den auf das Verhalten der Person gerichteten Erwartungen zu entsprechen (Fishbein & Ajzen, 1975). In dem aufgeführten Beispiel würde die Erwartung der Schulleitung, dass die Lehrkräfte der Schule Online-Lernplattformen für den Unterricht einsetzen (normative belief), zusammen mit der Motivation der betrachteten Lehrkraft, den Erwartungen der Schulleitung nachzukommen (motivation to comply), zu einer den Einsatz einer Online-Lernplattform befürwortenden subjektiven Norm führen.

Insgesamt lässt sich festhalten, dass die TRA eine Restrukturierung der drei Elemente von Einstellungen, wie sie bspw. im Drei-Komponenten-Modell von Rosenberg und Hovland (1969) zu finden sind, vornimmt (Ajzen, 1989). Kognitive und behaviorale Elemente werden hierbei nicht als inhärenter Teil von Einstellung betrachtet. Kognitive Elemente sind einer Einstellung in Form von Überzeugungen vorgelagert und beeinflussen die Einstellungsbildung ebenso wie die Herausbildung von Verhaltensnormen. Kognitive Prozesse im Sinne von Überzeugungen und die damit einhergehende Prämisse, dass einer Handlung immer kognitive Prozesse vorausgehen und alle Entscheidungen auf rationalen Überlegungen basieren, sind dabei ein zentrales Element der TRA. Kognitive Elemente beeinflussen das Verhalten jedoch nicht direkt, sondern lediglich durch ihre Wirkung auf Einstellungen und subjektive Norm. Diese werden nicht direkt, sondern erst durch die Differenzierung der behavioralen Komponente, mediiert über die Verhaltensintention, verhaltenswirksam. Die Ausführung eines Verhaltens kann sich wiederum im Sinne einer Feedbackschleife auf Überzeugungen auswirken und dadurch zukünftige Intentionen und Verhaltensweisen beeinflussen (Ajzen, 2014; Fishbein & Ajzen, 1975).

An dieser Stelle soll kurz auf die dem Modell zugrundeliegende Definition von Einstellung als affektives bzw. evaluatives Konstrukt eingegangen werden. Die synonyme Verwendung der Begriffe Affekt und Evaluation führt dazu, dass der evaluative Charakter von Einstellungen, der einen eigentlich kognitiven Vorgang beschreibt, mit der affektiven Komponenten von Einstellungen gleichgesetzt wird (Eagly & Chaiken, 1993; Zanna & Rempel, 1988). Ajzen und

Fishbein (2000) betonen zwar die Notwendigkeit einer klaren Differenzierung der Konstrukte, merken jedoch an, dass die Faktoren nicht immer klar unterschieden werden können und eine evaluative Reaktion durchaus von Stimmungen und Emotionen (affektive Elemente) beeinflusst werden kann. Das Fehlen einer eindeutigen Abgrenzung von Affekt und Evaluation und somit auch von affektiven und kognitiven Elementen von Einstellungen, stellt in der Literatur einen häufigen Kritikpunkt dar, nicht nur im Hinblick auf die TRA (Ajzen, 1989; Ajzen & Fishbein, 2000). Jüngere Arbeiten in diesem Kontext betrachten die beiden Konstrukte meist als konzeptionell unterschiedlich. Diesem Verständnis nach ist eine evaluative Komponente inhärenter Bestandteil aller – sowohl affektiver, kognitiver als auch behavioraler – Einstellungsreaktionen (Eagly & Chaiken, 2005). Evaluationen sind dabei ein intervenierender Zustand zwischen einem Einstellungsobjekt und einer durch diesen Stimulus hervorgerufenen evaluativen Reaktion. Affekte dagegen repräsentieren eine von drei Kategorien dieser evaluativen Einstellungsreaktionen (Eagly & Chaiken, 1993; 1998). Einstellungen, verstanden als die evaluative Bewertung eines Einstellungsobjektes, können dabei, wie bspw. Eagly und Chaiken (1998) anmerken, durchaus auch rein basierend auf kognitiven oder behavioralen Reaktionen zum Ausdruck kommen, ganz ohne affektive Reaktionen.

Andere Kritikpunkte der TRA betreffen bspw. die Betrachtung von Einstellung und subjektiver Norm als zwei separate Konstrukte. Kritische Stimmen sehen die Unterscheidung zwischen subjektiver Norm und Einstellung als eher willkürlich, da die beiden Konstrukte durchaus eng zusammenhängen können und subjektive Norm auch als ein Aspekt von Einstellung interpretiert werden kann (Eagly & Chaiken, 1993). Des Weiteren kommt Ajzen (1985) selbst zu dem Schluss, dass die TRA nicht in allen Situationen anwendbar ist, was zu einer Weiterentwicklung der Theorie führt, die im folgenden Kapitel beschrieben wird.

2.1.2.3. Theory of Planned Behavior (TPB)

Eine Weiterentwicklung der Theory of Reasoned Action stellt die Theory of Planned Behavior (TPB) dar (siehe Abbildung 4), in welcher Ajzen (1985, 1991) das ursprüngliche Modell überarbeitet und um eine weitere Komponente ergänzt. Wie bei der TRA, ist in der TPB die Verhaltensintention dem tatsächlichen Verhalten direkt vorgelagert (Ajzen, 1991). Die TRA sieht Intention jedoch eher als probabilistisches Konstrukt, im Sinne der subjektiven Wahrscheinlichkeit, ein bestimmtes Verhalten auszuführen (Fishbein & Ajzen, 1975), eine Konzeption, die bspw. von Warshaw und Davis (1985) eher als Verhaltenserwartung, weniger als Intention, charakterisiert wird. In der TPB beziehen sich Intentionen dagegen auf motivationale Faktoren

im Sinne der Bereitschaft bzw. dem Vorhaben einer Person, das fragliche Verhalten auszuführen bzw. sich darum zu bemühen, dieses auszuführen, und spiegeln somit die Verhaltensabsicht einer Person im Hinblick auf das Einstellungsobjekt wider (Ajzen, 1991). Auch in der TPB wird die Verhaltensintention wiederum durch die Einstellung gegenüber dem betrachteten Verhalten und durch die subjektive Norm beeinflusst (Ajzen, 1991). Da subjektive Norm und Einstellung sich jedoch als nicht hinreichend zur Erklärung des Verhaltens bzw. der Entstehung von Verhaltensintentionen erweisen, wurde im Zuge der Weiterentwicklung des Modells neben diesen beiden Einflussfaktoren die *wahrgenommene Verhaltenskontrolle (perceived behavioral control)* aufgenommen. Diese beeinflusst das Verhalten nicht nur indirekt, mediiert durch die Verhaltensintention, sondern unter bestimmten Voraussetzungen auch in direkter Weise (Ajzen, 1985, 1991).

Die Erweiterung der TRA wurde u. a. geleitet von der Erkenntnis, dass eine Verhaltensintention nur dann als adäquater Prädiktor des tatsächlichen Verhaltens gesehen werden kann, wenn das betrachtete Verhalten der *willentlichen Kontrolle (volitional control)* einer Person unterliegt, wodurch sich der Anwendungsbereich der TRA auf derartige Situationen begrenzt. In Situationen, in denen keine willentliche Kontrolle des betreffenden Verhaltens vorliegt, misslingt die in der TRA postulierte Vorhersage des Verhaltens aus der Verhaltensintention. Ein vielzitiertes Beispiel in diesem Kontext stellen Raucher*innen dar, die zwar, beeinflusst durch eine positive Einstellung und die Unterstützung des Familien- und Bekanntenkreises, eine positive Intention haben, mit dem Rauchen aufzuhören, dieses Vorhaben jedoch bedingt durch die Nikotinabhängigkeit nicht umsetzen können (Ajzen, 1985). Die Möglichkeit, ein Vorhaben im Sinne einer Verhaltensintention in ein tatsächliches Verhalten umzusetzen, ist demnach bedingt durch die tatsächliche Verhaltenskontrolle (*actual control*). Diese ist beeinflusst von verschiedenen internen Faktoren, wie individuellen Fähigkeiten, Informationen oder Willenskraft, aber auch von externen Faktoren, wie zeitlichen Vorgaben oder der Abhängigkeit von anderen Personen (Ajzen, 1985, 1991).

Da die tatsächliche Verhaltenskontrolle in der Realität nur schwer messbar ist, berücksichtigt die TPB die willentliche Kontrolle des Verhaltens durch die Integration des Konstrukts der wahrgenommenen Verhaltenskontrolle (Ajzen, 1991). Wahrgenommene Verhaltenskontrolle definiert sich dabei als die Wahrnehmung einer Person, wie leicht oder schwierig das betrachtete Verhalten auszuführen ist. Die wahrgenommene Verhaltenskontrolle wiederum wird bestimmt durch Überzeugungen hinsichtlich des Vorhandenseins oder der Abwesenheit von Ressourcen und Möglichkeiten, eine bestimmte Handlung auszuführen oder ein bestimmtes Ziel zu

erreichen, genannt *Kontrollüberzeugungen* (*control beliefs*). Kontrollüberzeugungen beziehen sich somit auf Faktoren, die ein bestimmtes Verhalten fördern oder behindern (Ajzen & Fishbein, 2005). Entstehen können Kontrollüberzeugungen bspw. durch die Bewertung eigener Erfahrungen, jedoch auch durch Informationen Dritter, wie Meinungen von Freunden oder durch sonstige erwartete Hindernisse (Ajzen & Madden, 1986; Ajzen, 1991). Sie beeinflussen die wahrgenommene Verhaltenskontrolle dabei in gleicher Weise wie normative und verhaltensbezogene Überzeugungen subjektive Norm bzw. Einstellung zum Verhalten beeinflussen (Ajzen & Madden, 1986; Eagly & Chaiken, 1993). Konkret ergibt sich die wahrgenommene Verhaltenskontrolle aus der Produktsumme der einzelnen Kontrollüberzeugungen im Sinne von Ressourcen und Möglichkeiten wie bspw. Zeit, eigene Fähigkeiten, finanzielle Mittel, und den daraus resultierenden wahrgenommenen fördernden bzw. hindernden Effekten (Ajzen, 1989, 2002). Eine Grundannahme der TPB ist, dass die wahrgenommene Kontrolle einer Person hinsichtlich des Verhaltens umso größer ist, je positiver diese Person ihre Ressourcen und Möglichkeiten einschätzt und je weniger Hindernisse gesehen werden (Ajzen & Madden, 1986). Eine Lehrkraft bspw., die ihre eigenen technischen Kenntnisse als unzureichend einschätzt (Kontrollüberzeugung) und dies als hinderlich für den Einsatz einer Online-Lernplattform erachtet (erschwerender Effekt), würde demnach eine geringe wahrgenommene Verhaltenskontrolle hinsichtlich der Nutzung entsprechender Tools herausbilden.

Zurückzuführen ist das Konstrukt der wahrgenommenen Verhaltenskontrolle auf Banduras (1977, 1982, 1991) Konzept der *Selbstwirksamkeit* bzw. *Selbstwirksamkeitserwartung*. Bandura (1977) definiert Selbstwirksamkeitserwartung als die Überzeugung, eine bestimmte Handlung erfolgreich ausführen zu können. Der Mechanismus der Selbstwirksamkeit beeinflusst dabei die Entscheidungen und Handlungen von Personen, einschließlich des Umfangs getätigter Investitionen zur Bewältigung bestimmter Situationen im Sinne von investierter Zeit und Aufwand, insbesondere beim Auftreten von Hindernissen. Menschen neigen dazu, Situationen zu vermeiden, die in ihrer Wahrnehmung mit den eigenen Kompetenzen schwer zu bewältigen sind. Auf der anderen Seite streben sie Situationen bzw. Aktivitäten an, deren Bewerkstelligung sie als problemlos ansehen. Das Verhalten einer Person wird demnach stark von ihrer Zuversicht in die eigenen Fähigkeiten beeinflusst (Bandura, 1977, 1991).

Im Kontext der TPB, in der Selbstwirksamkeit in dem Konzept der wahrgenommenen Verhaltenskontrolle repräsentiert ist, gibt es zwei Wirkmechanismen der subjektiv empfundenen Verhaltenskontrolle auf das Verhalten. Zum einen stellt diese, neben Einstellung zum Verhalten und subjektiver Norm, einen weiteren motivationalen Prädiktor der Verhaltensintention dar und

beeinflusst dadurch indirekt, als Mediatorvariable, das tatsächliche Verhalten (Ajzen & Madden, 1986). Dem liegt die Annahme zugrunde, dass Menschen dazu neigen, eine bestimmte Handlung anzustreben, wenn sie das Gefühl haben, dass diese ihrer Kontrolle unterliegt und sie sicher sind, diese Handlung mit ihren Fähigkeiten bewältigen zu können. Umgekehrt vermeiden sie ein Verhalten eher, wenn das Gefühl besteht, die Situation nicht kontrollieren und eine Aufgabe nicht bewältigen zu können (Eagly & Chaiken, 1993). Um das zuvor aufgeworfene Beispiel fortzuführen, würde eine geringe wahrgenommene Verhaltenskontrolle hinsichtlich des Einsatzes einer Online-Lernplattform aufgrund mangelnder technischer Kenntnisse der Nutzungsintention entgegenwirken, auch wenn die Lehrkraft grundsätzlich eine positive Einstellung zur Nutzung hat und eine befürwortende subjektive Norm vorliegt.

Im Fall einer adäquaten Repräsentation der tatsächlichen Kontrolle hinsichtlich der betrachteten Handlung durch die wahrgenommene Verhaltenskontrolle, kann die wahrgenommene Verhaltenskontrolle zum anderen als direkter Verhaltensprädiktor herangezogen werden (Ajzen & Fishbein, 2005). Streng genommen ist es jedoch in diesem Fall die tatsächliche Verhaltenskontrolle, die letztendlich direkt verhaltenswirksam wird, nicht die wahrgenommene Verhaltenskontrolle. Diese kann jedoch in dem Modell als hinreichende Näherung gesehen werden und ist im Gegensatz zur tatsächlichen Verhaltenskontrolle auch messbar (Ajzen & Madden, 1986). In dem betrachteten Beispiel würde die Einschätzung einer Lehrkraft der eigenen technischen Kompetenzen als unzureichend für den Einsatz digitaler Unterrichtstools deren Nutzung direkt negativ beeinflussen, wenn die Lehrkraft ihre Fähigkeiten korrekt bewertet. Schätzt eine Lehrkraft ihre technischen Fähigkeiten dagegen korrekterweise positiv ein, wirkt dies direkt verstärkend auf den Einsatz derartiger Instrumente für den Unterricht.

Insgesamt führen eine positive Einstellung und subjektive Norm gegenüber einem bestimmten Verhalten zusammen mit einer hohen wahrgenommenen Kontrolle im Hinblick auf das Verhalten zu einer starken Verhaltensintention und dies wiederum zu einem entsprechenden Verhalten. Im Falle einer realistischen Einschätzung der wahrgenommenen Verhaltenskontrolle wirkt diese zusätzlich direkt bestärkend auf die Ausführung der betrachteten Handlung (vgl. gestrichelter Pfeil in Abbildung 4) (Ajzen, 1991; Ajzen & Fishbein, 2000). Die Ausprägung der einzelnen Einflussfaktoren ist wiederum situativ abhängig, denn nicht alle wirken in jedem Szenario gleich stark. In manchen Situationen sind bspw. insbesondere behaviorale Überzeugungen und somit Einstellungen handlungsleitend, während subjektive Norm und wahrgenommene Verhaltenskontrolle kaum einen Einfluss haben. In anderen Situationen wird das Verhalten dagegen durch alle drei Einflussfaktoren annähernd gleichermaßen bestimmt (Maio & Haddock,

2010). Die Beeinflussung von Verhaltensweisen bzw. Verhaltensintentionen durch externe Faktoren beschränkt sich in der TPB, ebenso wie in der TRA, auf deren indirekten Einfluss über die Einflussnahme auf Überzeugungen (Ajzen & Albarracin, 2007; Fishbein & Ajzen, 2010). Auch wenn die TPB in den letzten Jahrzehnten in zahlreichen Arbeiten empirisch untersucht und ihre prädiktive Validität unter Beweis gestellt wurde, reichen die theoretischen Überlegungen der vorliegenden Arbeit noch darüber hinaus, weshalb im nächsten Abschnitt mit dem Technology Acceptance Model (TAM) eine weitere, auf TRA und TPB aufbauende, Theorie dargestellt wird.

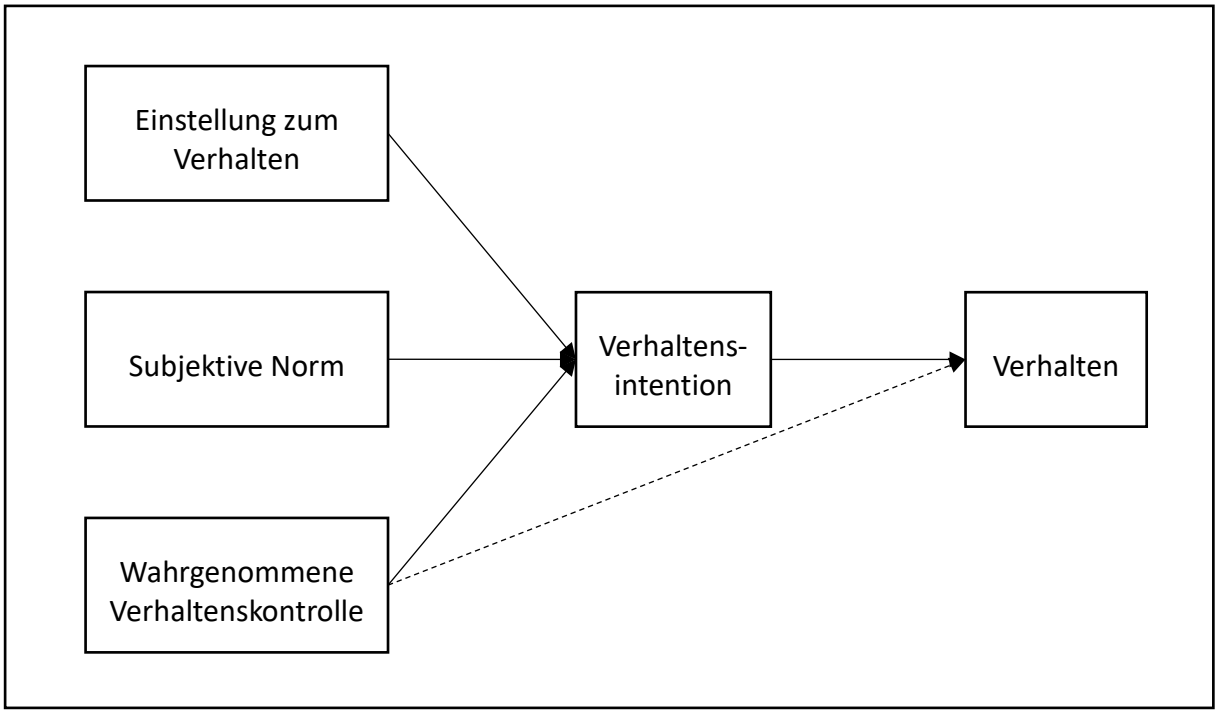


Abbildung 4: Theory of Planned Behavior (Darstellung nach Ajzen, 1991, S. 182)

2.1.2.4. Technology Acceptance Model (TAM)

Einen weiteren theoretischen Ansatz zur Erklärung von Nutzungsverhalten stellt das Technology Acceptance Model (TAM) nach F. D. Davis (1986) und F. D. Davis et al. (1989) dar (siehe Abbildung 5). Bei dem Modell handelt es sich um eine Theorie, deren primäres Anwendungsgebiet zwar im Bereich der Markt- und Konsumentenforschung zu verorten ist, jedoch durch die Nutzung von Einstellungstheorien als theoretische Grundlage eine sozialpsychologische Fundierung besitzt (F. D. Davis, 1993). Das TAM überträgt dabei die zentralen Annahmen der TRA in einen spezifischen betrieblichen Anwendungskontext und nutzt die TRA als theoretische Grundlage, die Akzeptanz von Informationssystemen in einem organisationsbezogenen

Umfeld zu erklären. Während die TRA somit einen sehr breiten Anwendungsbereich findet, wird das TAM, insbesondere in seinen Ursprüngen, speziell zur Erklärung der Akzeptanz und Nutzung von Computertechnologien genutzt. Ziel des Modells sind die Erklärung der Bestimmungsfaktoren der Akzeptanz technischer Systeme durch die Nutzer*innen und die daraus resultierende Erklärung des Nutzungsverhaltens (F. D. Davis et al., 1989).

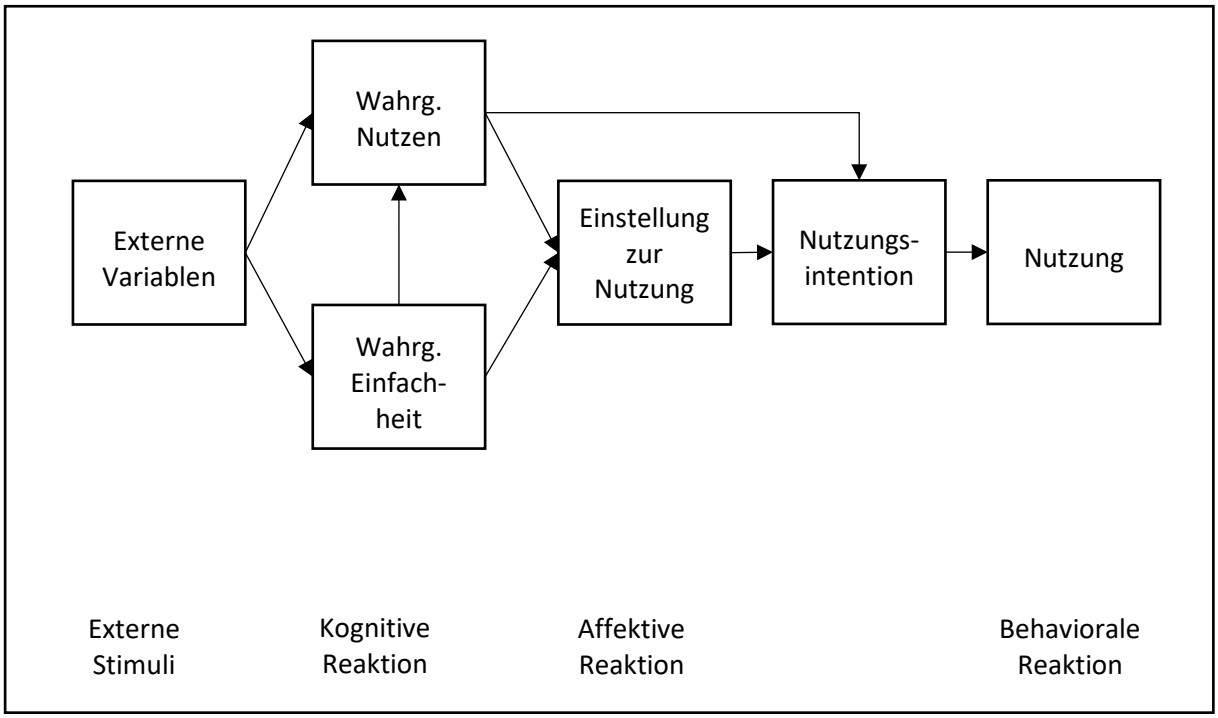


Abbildung 5: Technology Acceptance Model (Darstellung nach F. D. Davis et al., 1989, S. 985; F. D. Davis, 1993, S. 476)

Das TAM gebraucht, wie der Name bereits verrät, zum ersten Mal den Begriff Akzeptanz im Kontext von Einstellungstheorien, jedoch erfolgt über die Benennung des Modells hinaus keine Definition des Ausdrucks. Arbeiten zum TAM sprechen lediglich davon, die Akzeptanz, genauer die Akzeptanz der Nutzer*innen technischer Systeme, bzw. die Entstehung von Akzeptanz durch die Wahrnehmung der Nutzer*innen im Hinblick auf Einfachheit und Nutzen zu erklären (Venkatesh & Davis, 1996). Basierend auf den Modellannahmen lässt sich Akzeptanz im TAM als der Prozess der auf Wahrnehmung und Informationsverarbeitung bzw. -bewertung basierenden Einstellungsbildung und daraus entstehender Nutzungsintention bzw. Nutzung eines Systems interpretieren. Akzeptanz kann gemäß dieser Interpretation als Prozess verstanden werden, welcher in Abbildung 5 dargestellt wird. Auf Grundlage dieses aus dem TAM abgeleiteten Begriffsverständnis erfolgt in Kapitel 3 die für die vorliegende Arbeit gültige Definition von Akzeptanz.

In seinen Grundzügen folgt das TAM der Logik der TRA bzw. TPB, wonach die Konfrontation mit einem Einstellungsobjekt, im Falle des TAM einem technischen System, zunächst eine kognitive Reaktion herbeiführt, die zur Herausbildung einer Einstellung als Ausdruck einer affektiven Reaktion führt, woraufhin wiederum behaviorale Reaktionen folgen (F. D. Davis, 1986). Hinsichtlich der behavioralen Reaktionen ist beim TAM ebenso wie bei der TRA die *Nutzungsintention* (*behavioral intention to use*) direktes Antezedens der tatsächlichen Nutzung. Die Nutzungsintention wird wie in der TRA u. a. von der *Einstellung zur Nutzung* (*attitude toward using*) beeinflusst und darüber hinaus durch den wahrgenommenen Nutzen oder die *wahrgenommene Nützlichkeit* (*perceived usefulness*). Das TAM betrachtet Einstellungen, genauer die Einstellung zur Nutzung, ebenfalls angelehnt an die TRA, als affektive Reaktionen und definiert diese als affektive Bewertung, die eine Person mit der Nutzung eines Systems verbindet (F. D. Davis, 1993). Das Modell formuliert dabei die Annahme, dass Personen dann Verhaltensintentionen bilden, wenn sie einen positiven Affekt bzw. eine positive Emotion gegenüber dem betrachteten Verhalten besitzen (F. D. Davis et al., 1989). Wahrgenommene Nützlichkeit wird definiert als die von einer potenziellen Nutzerin bzw. einem potenziellen Nutzer subjektiv empfundene Wahrscheinlichkeit, dass die Nutzung einer bestimmten Anwendung bzw. eines Systems ihre oder seine Arbeitsleistung verbessert (F. D. Davis et al., 1989). Der direkte Einfluss einer Überzeugung in Form der wahrgenommenen Nützlichkeit auf die Nutzungsintention unterscheidet das TAM von der TRA. Diese direkte Verbindung ist geleitet von der Idee, dass Personen in einem betrieblichen Kontext, für welchen das Modell ursprünglich entwickelt wurde, tendenziell dann Verhaltensintentionen herausbilden, wenn sie davon ausgehen, dass ein Verhalten ihre Arbeit erleichtern wird, zunächst auch unabhängig davon, ob sie sonstige negative oder positive Gefühle dem Verhalten besitzen. Konkret ergibt sich die Nutzungsintention nun gemäß dem TAM aus der Summe des wahrgenommenen Nutzens und der Einstellung zur Nutzung (F. D. Davis et al., 1989).

Die Einstellung zur Nutzung wird im TAM u. a. durch den wahrgenommenen Nutzen bestimmt, welcher dadurch, neben seiner direkten Wirkung, zusätzlich einen indirekten Einfluss auf die Nutzungsintention ausübt. Neben dem wahrgenommenen Nutzen beeinflusst in dem Modell noch eine weitere Überzeugung in Form der *wahrgenommenen Einfachheit der Nutzung (perceived ease of use)* die Entstehung von Einstellungen (F. D. Davis et al., 1989). Die wahrgenommene Einfachheit beschreibt die Wahrnehmung einer potenziellen Nutzerin bzw. eines potenziellen Nutzers, wonach die Nutzung des betrachteten Systems ohne physischen und mentalen Aufwand erfolgt (F. D. Davis, 1986, 1989). Das Konstrukt der wahrgenommenen

Einfachheit lehnt sich dabei u. a. an das Konzept der Selbstwirksamkeit von Bandura (1977, 1982) an, welches auch der wahrgenommenen Verhaltenskontrolle der TPB zugrunde liegt. Je einfacher die Nutzung eines Systems ist, desto größer ist die wahrgenommene Selbstwirksamkeit im Hinblick auf das zur Nutzung erforderliche Verhalten. Dieser intrinsisch motivierte Aspekt der wahrgenommenen Einfachheit beeinflusst die Einstellung gegenüber der Nutzung (F. D. Davis et al., 1989). Darüber hinaus hat die wahrgenommene Einfachheit einen direkten Einfluss auf den wahrgenommenen Nutzen, da ein System bei gleichen Rahmenbedingungen umso nützlicher wahrgenommen wird, je einfacher es zu bedienen ist (F. D. Davis, 1986). Die dargestellten Zusammenhänge von Einfachheit, Nützlichkeit und Einstellung sollen in einem kurzen Beispiel verdeutlicht werden. Bspw. empfindet eine Lehrkraft den Einsatz einer bestimmten Online-Lernplattform als förderlich für den Unterricht, weil die Schüler*innen dadurch nicht nur fachliche Inhalte erlernen, sondern auch in ihrer Medienkompetenz geschult werden. Darüber hinaus nimmt die Lehrkraft die Anwendung des Tools als unkompliziert wahr. Die wahrgenommene Nützlichkeit zusammen mit der Wahrnehmung einer einfachen Anwendung befördert eine positive Einstellung zur Nutzung, welche wiederum die Intention, das Tool im eigenen Unterricht einzusetzen, begünstigt. Darüber hinaus besteht bei einem Unterrichtstool, das als nützlich empfunden wird, eine höhere Nutzungsintention als bei einem Tool, das von einer Lehrkraft als nicht förderlich wahrgenommen wird. Im Hinblick auf die Einfachheit wird ein Instrument, das einfach zu bedienen scheint, auch als nützlicher empfunden als ein kompliziert einzusetzendes Tool.

Insgesamt ergibt sich der wahrgenommene Nutzen aus der Summe der wahrgenommenen Einfachheit und externen Variablen. Bei den externen Variablen handelt es sich um Eigenschaften des betrachteten Systems, wie Designelemente, aber auch bspw. weitere Unterstützungsangebote im Zusammenhang mit dem System. Auch die Wahrnehmung der Einfachheit wird im TAM durch diese externen Variablen beeinflusst (F. D. Davis et al., 1989). Externe Variablen haben somit keinen direkten Einfluss auf die Einstellungsbildung oder die Nutzung, sondern wirken lediglich über die wahrgenommene Einfachheit und die Nutzenwahrnehmung (F. D. Davis, 1986; F. D. Davis et al., 1989). Das folgende Beispiel soll die Wirkungsweise externer Faktoren im Rahmen des TAM illustrieren. Die Gestaltung einer Online-Lernplattform beeinflusst bspw. deren Wahrnehmung bzgl. einer einfachen Nutzung und der Nützlichkeit. Bietet ein solches Tool etwa die Möglichkeit, mit wenigen Klicks eine direkte Rückmeldung für die Schüler*innen zu erzeugen, im Gegensatz zu einem Tool, bei dem diese durch die Lehrkraft selbst erstellt werden muss, wird das erste Tool als unkompliziert und einfacher einsetzbar

empfunden. Durch die wahrgenommene Einfachheit der Bedienung steigt aufgrund der daraus resultierenden angenommenen Zeit- bzw. Arbeitsersparnis der wahrgenommene Nutzen. Der direkte Einfluss der Beschaffenheit eines Systems auf den wahrgenommenen Nutzen zeigt sich bspw. bei der Betrachtung zweier vergleichbar leicht zu bedienender digitaler Tools, die beide eine direkte Rückmeldung für Schüler*innen bieten. Ist diese bei einer Variante performanter oder gibt es eine übersichtlichere Darstellungsweise, wirkt dies direkt positiv auf den wahrgenommenen Nutzen.

Im TAM repräsentieren wahrgenommene Einfachheit und wahrgenommener Nutzen die kognitiven Reaktionen auf die Konfrontation mit dem betrachteten Einstellungsobjekt, analog zu den verhaltensbezogenen Überzeugungen der TRA. Wie in der TRA betrachtet das TAM wahrgenommene Einfachheit und wahrgenommen Nutzen zwar als eigenständige Konstrukte, definiert jedoch anders als die TRA durch den angenommenen Einfluss der Einfachheit auf die Nutzenwahrnehmung eine kausale Beziehung zwischen diesen verschiedenen Überzeugungen. Eine solche Kausalbeziehung existiert in der TRA und TPB nicht, dort werden die Überzeugungen lediglich gewichtet und zu einem einzelnen Konstrukt aufaddiert, mögliche Beziehungen zwischen den Überzeugungen bleiben unberücksichtigt (F. D. Davis, 1986; F. D. Davis et al., 1989). Die Entstehung der Einstellung durch Überzeugungen, hier wahrgenommener Nutzen und wahrgenommene Einfachheit, im TAM ist dagegen durchaus vergleichbar mit der TRA. Im TAM wird der Einfluss jedoch statistisch durch Regressionen bzw. vergleichbare Methoden wie Strukturgleichungsmodelle ermittelt und nicht durch die Produktsumme aus Überzeugungen und deren Bewertungen wie in der TRA (F. D. Davis et al., 1989).

Hinsichtlich der Vorhersagekraft der beiden Konstrukte auf die Nutzung bzw. vorgelagert die Einstellung gegenüber der Nutzung zeigen Untersuchungen, dass der wahrgenommene Nutzen einen deutlich stärkeren Einfluss ausübt als die wahrgenommene Einfachheit und sogar direkt die Nutzung eines Systems bestimmen kann. Dieser im TAM postulierte und empirisch belegte direkte Effekt einer Wahrnehmung in Form des wahrgenommenen Nutzens auf eine Handlung unterscheidet das Modell von der TRA, in welcher Wahrnehmungen bzw. Überzeugungen zum Einstellungsobjekt nur gemittelt über die Einstellung verhaltenswirksam werden, und unterstreicht die Bedeutung des wahrgenommenen Nutzens zur Erklärung von Verhalten (F. D. Davis, 1993). Die Wirkung der wahrgenommenen Einfachheit zeigt sich dagegen nur durch einen starken Effekt auf die Nutzenwahrnehmung und den dadurch entstehenden indirekten Effekt auf die Einstellung (F. D. Davis, 1993; Venkatesh & Davis, 2000). Die Bedeutsamkeit der Nutzenwahrnehmung zur Vorhersage von Verhalten bzw. Verhaltensintentionen wurde in den

letzten Jahrzenten im Kontext des TAM vielfach empirisch belegt (F. D. Davis, Bagozzi & Warshaw, 1992; Ho Cheong & Park, 2005; Kulviwat, Bruner II, Kumar, Nasco & Clark, 2007; Venkatesh & Davis, 2000; Yang & Yoo, 2004; Zhao, Fang & Jin, 2018). Mehrere Metaanalysen bestätigen einen durchweg starken Einfluss der wahrgenommenen Nützlichkeit auf Nutzung oder Nutzungsintention, während die wahrgenommene Einfachheit meist primär über ihren Einfluss auf die Nutzenwahrnehmung wirkt (Chismar & Wiley-Patton, 2003; King & He, 2006; Ma & Liu, 2004). Die Nutzungsintention erweist sich dabei, wie im ursprünglichen TAM vorhergesagt, in der Metaanalyse von Turner, Kitchenham, Brereton, Charters und Budgen (2010) als valider Prädiktor der tatsächlichen Nutzung.

Insgesamt unterscheidet das TAM, ebenso wie die TRA bzw. TPB, zwischen kognitiven, affektiven und behavioralen Reaktionen auf ein Einstellungsobjekt. Ähnlich wie in der TRA führen kognitive Prozesse in Form von Überzeugungen bzw. Wahrnehmung, im TAM repräsentiert durch wahrgenommenen Nutzen und wahrgenommene Einfachheit der Nutzung, zur Einstellungsbildung, einer primär affektiven Reaktion. Die Einstellung gilt, ebenso wie in der TRA, auch im TAM als direkter Verhaltensprädiktor bzw. wirkt direkt auf die Nutzungsintention, die sowohl gemäß TAM als auch der TRA das Verhalten vorhersagt. Der direkte Einfluss eines kognitiven Elements, in Form des wahrgenommenen Nutzens, auf eine behaviorale Komponente, hier die Nutzungsintention, unterscheidet das TAM von der TRA, in der ein solch direkter Einfluss kognitiver Elemente auf Verhaltensaspekte nicht existiert. Externe Variablen jedoch wirken auch im TAM, vergleichbar mit der TRA, lediglich indirekt, über ihren Einfluss auf Wahrnehmungsfaktoren, auf Einstellung und Verhalten. Grundsätzlich kann das TAM, welches für einen betrieblichen Anwendungsbereich, genauer zur Untersuchung der Akzeptanz innovativer technologischer Systeme, entwickelt wurde, nicht einfach unverändert in einen pädagogischen Kontext übernommen werden, sondern bedarf entsprechender Anpassungen. Auch F. D. Davis (1993) und Venkatesh und Davis (2000) merken an, dass je nach Anwendungsbereich ggf. weitere Variablen in das Modell aufgenommen werden müssen. Jedoch erweisen sich gerade die Kernelemente des TAM, also die Wirkung der Einflussfaktoren wahrgenommene Einfachheit und wahrgenommener Nutzen, als wichtigste Prädiktoren von Verhalten bzw. Verhaltensintention (King & He, 2006).

2.2. Vergleichsarbeiten – VERA

Das folgende Unterkapitel 2.2.1 dient der Erläuterung der Konzeption, Ziele und Funktionen von Vergleichsarbeiten und deren Einordnung in den übergeordneten Kontext der Gesamtstrategie zur *Neuen Steuerung* im deutschen Bildungswesen. Die Darstellungen beschränken sich dabei auf die zum Verständnis notwendigen Ausführungen. Weiterführende Informationen und konzeptionelle Hintergründe zu den Vergleichsarbeiten sowie deren Einordnung in die Gesamtstrategie zum Bildungsmonitoring (KMK, 2016) wurden bereits vielfach ausführlich in der Literatur behandelt. Bspw. sei hier auf Stanat, Pant, Pöhlmann und Kuhl (2013) und Altrichter et al. (2016) verwiesen. Im darauffolgenden Kapitel 2.2.2 werden zentrale Erkenntnisse aus Forschungsarbeiten im Kontext von VERA skizziert, mit Fokus auf Erkenntnissen zur Perspektive von Lehrkräften, besonders unter dem Aspekt der Untersuchung von Akzeptanz.

2.2.1. Instrumente der Neuen Steuerung: Konzeption, Ziele und Funktionen von Bildungsstandards und Vergleichsarbeiten

2.2.1.1. Bildungsstandards

Mit dem Beschluss der Kultusministerkonferenz zur Implementierung einer Gesamtstrategie zur Qualitätssicherung und -entwicklung in deutschen Schulen erfolgte im Jahr 2002 die Einführung bundesweit geltender Bildungsstandards (KMK, 2004b). Auslöser der Implementierung der Bildungsstandards war nicht zuletzt das vergleichsweise erschreckend schlechte Abschneiden deutscher Schüler*innen in internationalen Vergleichsstudien, wie PISA oder zuvor bereits TIMSS (Maaz et al., 2019), bei denen auch bundesintern teils deutliche Leistungsunterschiede zwischen Bundesländern hervortraten (Artelt, Schneider & Schiefele, 2002). Diese Befunde legten die Notwendigkeit zur Reform des deutschen Bildungssystems nahe (Zimmer-Müller, Hosenfeld & Koch, 2014). Entworfen wurde ein System der Neuen Steuerung mit dem Ziel einer stärkeren Standard-, Evidenz- und Outputorientierung des Bildungssystems. Steuerung auf Ebene der Schulen ist dabei als die Gestaltung von Lehr-Lern-Prozessen zu verstehen, welche das Ziel verfolgt die Kompetenzentwicklung von Schüler*innen zu fördern und Differenzen zu erwarteten Leistungen zu verringern bzw. Differenzen zum Ausgangszustand zu vergrößern (Diemer & Kuper, 2011). Eingebettet in dieses Steuerungssystem wurden im Zuge der Reformbemühungen zu Beginn der 2000er Jahre u. a. Bildungsstandards und bundesweite

Vergleichsarbeiten als zentrale Steuerungselemente etabliert (Maag Merki, 2016; Maaz et al., 2019).

Bildungsstandards legen fachbezogene Kompetenzniveaus fest, welche Schüler*innen in einer bestimmten Klassenstufe erreicht haben sollten (KMK, 2016), das heißt, sie formulieren Erwartungen an fachliches Lernen im Kontext allgemeiner Bildungsziele (Klieme et al., 2003). Mit der Implementierung einer verstärkten Kompetenzorientierung für Bildungsprozesse ging das Ziel einher, den Erwerb von vernetztem Wissen und einer damit verbundenen Befähigung zu flexibler Problembewältigung bei Schüler*innen zu fördern (Stanat et al., 2013). Die entsprechenden Kompetenzen bzw. Kompetenzanforderungen werden in den Bildungsstandards so konkret beschrieben, dass diese in Testaufgaben übersetzt und durch spezielle Testverfahren erfasst werden können (Klieme et al., 2003). Die Konzeption derartiger standardisierter Testformate zur Untersuchung erreichter Kompetenzen beschreibt eine der Kernfunktionen von Bildungsstandards, die *Überprüfungsfunktion* (Maaz et al., 2019). Des Weiteren benennen Maaz et al. (2019) im Rückgriff auf die ausführliche Beschreibung der Bildungsstandards durch Klieme et al. (2003) drei weitere zentrale Funktionen von Bildungsstandards: Zunächst haben diese eine *Orientierungsfunktion*, indem sie einen Referenzrahmen für Schulqualität auf allen Systemebenen und für alle betroffenen Akteur*innen bereitstellen. Durch die Definition dessen, was unter einer bestimmten Kompetenz zu verstehen ist, erfüllen sie darüber hinaus eine *Klärungsfunktion*. Aus der bereits angesprochenen Überprüfung der Kompetenzen lassen sich wiederum Ansatzpunkte zur Unterrichtsentwicklung und -veränderung ableiten, wodurch den Bildungsstandards zuletzt eine *Entwicklungsfunktion* zukommt. Zusätzlich ergibt sich aus der Kompetenzstandüberprüfung, neben einer Entwicklungsfunktion auf Schulebene, auf der Ebene der operativen Schulaufsicht eine *Controllingfunktion,* wenn die erhobenen Daten dieser zur Rechenschaftslegung und Zielvereinbarung bereitgestellt werden (F. Thiel et al., 2019).

Bildungsstandards ermöglichen somit ein klassen-, schul- und länderübergreifendes Monitoring der Leistungsfähigkeit des Schulsystems (Maaz et al., 2019) und können vor dem Hintergrund der neuen Steuerungslogik des Bildungssystems als Inputfaktoren verstanden werden, welche die Zieldimensionen im Sinne des zu einem bestimmten Zeitpunkt zu erreichenden Kompetenzniveaus festlegen. Auf der anderen Seite eines derartigen Steuerungsmodells stellen Vergleichsarbeiten (VERA) – ebenso wie bspw. zentrale Abschlussprüfungen – ein Output prüfendes Element dar, indem sie definierte Leistungen und Kompetenzen von Schüler*innen am Ende einer festgelegten Lernperiode erfassen (Maag Merki, 2016). Durch die darauf basierende Bereitstellung einer zusätzlichen Datengrundlage können Vergleichsarbeiten als ein Instrument

zur weiteren Implementation der Bildungsstandards auf Schulebene, sowie als Instrument der Unterrichtsentwicklung genutzt werden (Kühle & Peek, 2007). Inwiefern zentrale Vergleichsarbeiten ihre Funktion als Steuerungsinstrument erfüllen, hängt dabei von der Nutzung der Ergebnisse in den Schulen ab. Zunächst handelt es sich bei VERA lediglich um ein Testinstrument zur Feststellung der Leistungen von Schüler*innen. Erst durch eine Rückmeldung der Daten an schulische Akteur*innen und eine aktive Auseinandersetzung mit diesen Daten und deren Nutzung zur Gestaltung von Lehr-Lern-Prozessen entfaltet das Instrument auf schulischer Ebene seine Steuerungswirkung (Diemer & Kuper, 2011; Hartung-Beck & Diemer, 2009). Mit den Gelingensbedingungen auf Schulebene beschäftigt sich Kapitel 2.2.2 durch die Aufarbeitung der Sichtweise von Lehrkräften auf Vergleichsarbeiten. Zunächst werden jedoch in den folgenden Abschnitten die Grundlagen von VERA erläutert.

2.2.1.2. Implementierung und Ziele von Vergleichsarbeiten

Im Zuge der Einführung der bundesweit einheitlichen Bildungsstandards wurden die Vergleichsarbeiten bereits in den Jahren 2002 und 2003 initiiert und erstmals 2003 zunächst zu Beginn der vierten Klasse im Fach Mathematik in mehreren deutschen Bundesländern durchgeführt. Seit 2008 erfolgt mit VERA3 eine deutschlandweit flächendeckende Testung in der dritten Jahrgangsstufe, seit 2010 wird auch in den achten Klassen bundesweit getestet (Richter, Böhme, Becker, Pant & Stanat, 2014; Richter, 2016). Seit dem Schuljahr 2019/2020 nimmt jedoch das Land Niedersachsen nicht mehr an den Vergleichsarbeiten teil, nachdem die Entscheidung zur Teilnahme bereits im Jahr zuvor den einzelnen Lehrkräften übertragen wurde (Niedersächsisches Kultusministerium, 2019).[3] In den verbliebenen 15 Bundesländern werden in der Primarstufe die Fächer Deutsch und Mathematik, in der Sekundarstufe I, genauer in der achten Klasse, zusätzlich die Fremdsprachen Englisch und Französisch geprüft (Richter & Böhme, 2014). Die Aufgaben der Vergleichsarbeiten bilden daher die Bildungsstandards ab, die Ende der 4. Klasse (VERA3) bzw. Ende der 10. Klasse (VERA8) erreicht werden sollten (Maaz et al., 2019). Kompetenzstufenmodelle veranschaulichen dementsprechend in den jeweiligen Fächern die verschiedenen erreichbaren Niveaus der in den Bildungsstandards beschriebenen Kompetenzen, wobei Kompetenzstufe 1 den unteren Mindeststandard beschreibt und Stufe 5 den Optimalstandard. Auch die Rückmeldung im Rahmen der Vergleichsarbeiten

[3] Der Ausstieg des Landes Niedersachsen aus der Durchführung der Vergleichsarbeiten ist für den empirischen Teil dieser Arbeit von Relevanz, da sich dadurch die Datengrundlage zwischen den untersuchten Jahren 2018 und 2019 verändert. Entsprechend erfolgt bei der Beschreibung der Stichprobe ein weiterer diesbezüglicher Hinweis (siehe Kapitel 4.2.1. und 4.2.5).

erfolgt i. d. R. u. a. unter Bezugnahme auf dieses Kompetenzstufenmodell (Vettorazzi, Emmrich & Fuchs, 2017).

Die Vergleichsarbeiten werden zwar i. d. R. jährlich durchgeführt, haben aber dennoch den Charakter einer Querschnittserhebung, da die Durchführung pro Kohorte einmalig in der festgelegten Klassenstufe erfolgt. Die rückgemeldeten Daten liefern somit eine Momentaufnahme der Leistungsverteilung der getesteten Schüler*innen-Kohorte zum Testzeitpunkt (Kuper, 2008). Die VERA-Rückmeldungen können den Lehrkräften über das Referenzsystem der jeweiligen Lerngruppe hinaus als substanzielle Ergänzung ihrer diagnostischen Informationen dienen. Da sich die Vergleichsarbeiten nicht auf den direkt vorangegangenen Unterrichtsstoff beziehen und aufgrund ihrer Eigenschaft als diagnostisches Instrument, ist eine Benotung nicht zulässig (Vettorazzi et al., 2017).

Zwischen den einzelnen Bundesländern gibt es einige Unterschiede bei der Implementierung der Vergleichsarbeiten. U. a. variieren die verpflichtenden Testfächer bzw. Testdomänen je nach Land, zusätzlich aber u. U. auch nach Durchführungsjahr (Tarkian, Maritzen, Eckert & Thiel, 2019; Wurster, Bach et al., 2016). Während sich darüber hinaus in den meisten Ländern der Name VERA für die Vergleichsarbeiten durchgesetzt hat, sind diese in einigen Bundesländern unter anderem Namen verbreitet. In Hamburg unter dem Namen „KERMIT – Kompetenzen ermitteln“ bekannt, sind in Sachsen und Thüringen „Kompetenztests“, in Hessen und Nordrhein-Westfalen „Lernstandserhebungen“ die geläufigen Begrifflichkeiten. Alle Bezeichnungen benennen jedoch dieselbe Testung (Sälzer, 2016; Tarkian et al., 2019). In vorliegender Arbeit werden die Begriffe Vergleichsarbeiten, VERA und Lernstandserhebungen im Weiteren synonym verwendet. Unterschiede zwischen den Ländern bestehen auch hinsichtlich der Entscheidungsbefugnisse verschiedener Akteur*innen wie Schulaufsicht, Schulleitungen und weiterer schulischer Gremien. Ebenso unterscheiden sich die Nutzung der VERA-Rückmeldungen, bspw. im Hinblick auf verpflichtende Elternrückmeldungen und die Ausgestaltung verschiedener Anschlussaktivitäten, wie die Bereitstellung zusätzlicher Unterstützungsangebote (F. Thiel et al., 2019). Die jeweiligen Bundesländer sind auch für die Koordination der Durchführung von Vergleichsarbeiten verantwortlich. In den Ländern sorgen daher Landesinstitute und Qualitätsagenturen sowie auswertende Einrichtungen an Hochschulen, wie bspw. das Zentrum für Empirische Pädagogische Forschung (zepf) der Rheinland-Pfälzisch Technischen Universität Kaiserslautern-Landau (RPTU), ehemals Universität Koblenz-Landau, oder auch die entsprechenden Fachbereiche der Ministerien, für den Druck und die Verteilung der Testmaterialien und koordinieren die Testdurchführung und Auswertung sowie die Rückmeldung der

Ergebnisse an die Schulen. Die Entwicklung der Testaufgaben und deren Pilotierung obliegt seit dem Jahr 2008 ebenso dem Institut für Qualitätsentwicklung im Bildungswesen (IQB) wie die Skalierung der Aufgaben sowie die Erstellung der Testhefte und die Bereitstellung didaktischer Materialien zur Weiterarbeit in den Schulen (Sälzer, 2016).

Trotz vieler Unterschiede in der Implementierung der Vergleichsarbeiten zwischen den Bundesländern lässt sich die Entwicklungsfunktion der Vergleichsarbeiten als übergeordnetes Ziel identifizieren, auf das sich alle Länder verständigen können (F. Thiel et al., 2019). Ramsteck und Maier (2015) unterscheiden weitergehend verschiedene Zieldimensionen der Vergleichsarbeiten je nach Systemebene. Entwicklung findet demnach auf mehreren schulischen Ebenen statt. Auf Ebene der Lehrkräfte (Mikroebene) wird durch Hinweise zur Unterrichtsreflexion und Förderbedarfe eine Unterrichtsentwicklung angestrebt, während auf der Mesoebene der Schulleitungen durch datenbasierte schulische Selbstevaluation und die Stärkung der Gestaltungsautonomie von Schulen, die Schulentwicklung gefördert wird. Das Potenzial zur Schulentwicklung liegt hier insbesondere im kollegialen Austausch und einer gemeinsamen Diskussion der Ergebnisse.

Darüber hinaus steht auf der Makroebene letztendlich für Schulaufsicht und Bildungspolitik das Bildungsmonitoring im Vordergrund, welches durch den empirischen Abgleich von Mindeststandards (Rechenschaftslegung) und die Sicherung der Leistungsfähigkeit des Bildungssystems realisiert wird. Das gesamte Schulsystem betreffend, soll des Weiteren eine ebenenübergreifende Kommunikation zwischen den Akteur*innen gefördert werden (Skejic et al., 2015).

2.2.1.3. Ablauf der Vergleichsarbeiten

Während die Bundesländer als übergeordnete Instanz die Testadministration in ihrem jeweiligen Land zu verantworten haben (Sälzer, 2016), obliegen Testdurchführung und Auswertung auf Schulebene i. d. R. den jeweiligen Fachlehrkräften (Tarkian et al., 2019; Wurster, Bach et al., 2016). Die Auswertung erfolgt zwar nicht frei nach der Einschätzung der Lehrkräfte, sondern anhand einer standardisierten Auswertungsanleitung, dennoch bietet dieses Vorgehen durchaus Angriffspunkte, die Güte des Testinstrumentes insbesondere mit Blick auf die Objektivität in Frage zu stellen (Vogel, 2020). Gerade die u. a. durch die Vergleichsarbeiten intendierten kriterialen und sozialen Vergleiche sind demnach nur dann sinnvoll interpretierbar,

wenn gewisse Mindeststandards der Objektivität und entsprechend der Reliabilität und Validität gegeben sind (Spoden, Fleischer & Leutner, 2014).

Durch die externe Entwicklung der Testaufgaben entsteht den Lehrkräften zwar zunächst keine zusätzliche Belastung, die Durchführung und Auswertung sowie die Eingabe der Ergebnisse erfordert jedoch je nach Fach und Testheft durchaus einen zeitlichen Aufwand, welcher häufig als weiterer Kritikpunkt an Vergleichsarbeiten angemerkt wird (Richter, 2016). Nach der Durchführung der Vergleichsarbeiten, gefolgt von Auswertung und Ergebniseingabe, ist die Aufgabe der Lehrkräfte zunächst zwar erfüllt, jedoch erhalten sie nun von den jeweiligen auswertenden Einrichtungen eine Rückmeldung über die Kompetenzstände ihrer Schüler*innen, die die Lehrkräfte wiederum dazu anregen sollen, sich vertieft mit ihren Schüler*innen und ihrem Unterricht auseinanderzusetzen (Zimmer-Müller et al., 2014). Die Inhalte der Rückmeldungen werden im folgenden Abschnitt näher beschrieben.

2.2.1.4. VERA-Rückmeldungen

Die zentralen Rückmeldungen informieren über den aktuellen Lernstand der Schüler*innen hinsichtlich der Bildungsstandards im Sinne einer Momentaufnahme (G. Fuchs & Brunner, 2017; Stanat et al., 2013). Dabei sollen sie die fachdidaktische Diskussion und Kooperation in den Schulen befeuern (Krelle, 2015), schlagen jedoch keine konkreten Maßnahmen vor, sondern weisen lediglich auf Schwächen und Stärken in bestimmten Bereichen hin und geben Ansatzpunkte zur Ableitung eigener Veränderungsmaßnahmen (Isaac, Halt, Hosenfeld, Helmke & Groß Ophoff, 2006). Konkret liefern Vergleichsarbeiten u. a. diagnostische Informationen auf Individualebene und ermöglichen durch die Auskunft über Kompetenzstände der Schüler*innen eine individuell passgenaue Förderung der jeweiligen Kompetenzen (Gasteiger & Krelle, 2018; Maier et al., 2012).

Lehrkräften und Schulleitungen soll VERA somit als eine Art Frühwarnsystem dienen, um mögliche Kompetenzdefizite der Schüler*innen rechtzeitig auszumachen und entsprechende Fördermaßnahmen umsetzen zu können (Stanat et al., 2013). Auch die Erkenntnisse der Forschungsarbeit von Graf, Harych, Wendt, Emmrich und Brunner (2016) untermauern die Funktion der Vergleichsarbeiten als ein Instrument zur frühzeitigen Diagnose etwaiger schulischer Defizite in bestimmten Kompetenzbereichen und der damit einhergehenden Möglichkeit, rechtzeitig mit entsprechenden Maßnahmen entgegenzusteuern. Die repräsentative Längsschnittstudie an Berliner Gymnasien zur Prognosegüte von VERA8 zeigt substanzielle Korrelationen

zwischen VERA-Ergebnissen und sowohl den Ergebnissen der zentralen Abschlussprüfungen zum Mittleren Schulabschluss Ende der 10. Klasse als auch zu den korrespondierenden Jahrgangsnoten. Darüber hinaus erweisen sich die VERA-Ergebnisse im Fach Mathematik als Prädiktoren eines linearen bzw. nicht-linearen Bildungsverlaufs. Hinweise zum prognostischen Mehrwert von Vergleichsarbeiten finden sich auch bei G. Fuchs und Brunner (2017), die anhand des Abschneidens bei Vergleichsarbeiten Prognosen über den weiteren schulischen Erfolg von Schüler*innen ableiten. Die Autor*innen zeigen in ihrer Studie mit Brandenburger Schüler*innen, dass bildungsstandardbasierte Tests einen prognostischen Mehrwert haben. Insbesondere zeigen diese Tests das Potenzial, zukünftige Testleistungen und Schulnoten in den Fächern Mathematik und Deutsch vorherzusagen. Darüber hinaus weisen die Tests, wenn sie in beiden Fächern zusammengenommen werden, eine gute Prognosegüte für eine spätere Gymnasialempfehlung auf.

Trotz vieler positiver Befunde ist die Nutzung der Testrückmeldungen zur Individualdiagnostik nicht unumstritten. Bspw. merken Leutner, Fleischer, Spoden und Wirth (2008) an, dass eine individualdiagnostische Nutzung aufgrund der begrenzten Anzahl an Items nur eingeschränkt möglich ist. Auch Krelle (2015) verweist darauf, dass die Vergleichsarbeiten keine adäquaten Individualtests darstellen und die Ergebnisrückmeldungen nur auf Klassenebene valide analysierbar und interpretierbar sind.

Generell erfolgen die Rückmeldungen an die verschiedenen schulischen Akteur*innen (z. B. Schulleitungen, Lehrkräfte, Fachgruppen) auf unterschiedlichen Systemebenen: auf Ebene der einzelnen Klassen bzw. Lerngruppen, auf Schulebene sowie auf Landesebene (Zuber & Altrichter, 2018). Inhaltlich bilden die Rückmeldungen verschiedene Abstraktionsebenen ab: Zunächst erfolgt eine Auswertung der Lösungshäufigkeiten je Aufgabe, um den Lehrkräften eine Auseinandersetzung mit den Aufgaben zu ermöglichen. Darüber hinaus werden die im Test erreichten Kompetenzstufen zurückgemeldet, welche verständliche Vergleiche der Leistungen der eigenen Klasse mit anderen Gruppen wie der Schule oder dem gesamten Land ermöglichen und für eine sich bestenfalls anschließende Unterrichtsentwicklung besonders wichtig sind (Zimmer-Müller et al., 2014). Durch die Einordnung der Leistungen der eigenen Schüler*innen in die durch die Bildungsstandards definierten Kompetenzen ermöglichen die Vergleichsarbeiten zum einen eine kriteriale Rückmeldung. Durch den Abgleich der Kompetenzen der eigenen Klasse mit anderen Gruppen von Schüler*innen (z. B. Parallelklasse, -schule oder Bundesland) ergeben sich zum anderen soziale Vergleichsmöglichkeiten (Leutner et al., 2008).

In einigen Bundesländern erhalten die Lehrkräfte bei VERA3 eine zusätzliche soziale Referenz, den sogenannten *Fairen Vergleich*. Dies ist eine Rückmeldung im Vergleich zu definierten Vergleichsgruppen, basierend auf bestimmten Kontextfaktoren wie bspw. dem sozio-ökonomischen Status der Eltern oder dem Migrationshintergrund der Schüler*innen. Dies dient dazu, der sozialen Zusammensetzung von Klassen Rechnung zu tragen (Zimmer-Müller et al., 2014; Zuber & Altrichter, 2018). Diese Form der Rückmeldung ermöglicht es, Vergleichswerte zwischen Klassen und Schulen mit ähnlichen Rahmenbedingungen und Voraussetzungen abzubilden (Krelle, 2015). Fiege, Reuther und Nachtigall (2011) weisen dahingehend in ihrem Forschungsbeitrag jedoch darauf hin, dass eine kausale Interpretation mit Vorsicht behandelt werden muss und eigentlich eher von „faireren" Vergleichen zu sprechen sei, betonen aber dennoch die Bedeutung derartiger Adjustierungen der Testergebnisse, da die Berücksichtigung diverser Kontextvariablen durchaus eine validere Analyse von Unterrichtseffekten ermöglicht.

2.2.1.5. Abgrenzung zu anderen Leistungstests

Vergleichsarbeiten unterscheiden sich nicht nur von schulinternen Leistungsmessungen wie bspw. Klassenarbeiten, bei welchen die Entwicklung der Aufgaben sowie deren Auswertung i. d. R. in der Hand der Fachlehrkraft liegen und sich die Testinhalte auf den vorangegangenen Unterricht beziehen (Bonsen, Büchter & Peek, 2006). U. a. das Vorgehen bei der Durchführung der Vergleichsarbeiten unterscheidet diese auch von anderen internationalen und nationalen System-Monitoring-Instrumenten wie bspw. PISA oder dem IQB-Bildungstrend, vormals IQB-Ländervergleich, bei denen eine Testdurchführung durch externe Testleiter*innen erfolgt (Sälzer, 2016). Die Verfahrensweise einer schulintern gesteuerten Durchführung führt zwar bei den Vergleichsarbeiten zu einem geringeren Grad der Standardisierung (Richter et al., 2014), ermöglicht jedoch eine vergleichsweise zeitnah zum Test erfolgende Ergebnisrückmeldung an die Lehrkräfte und Schulen, was bei anderen Monitoring-Instrumenten u. U. Monate bis Jahre in Anspruch nehmen kann (KMK, 2016; Sälzer, 2016). Dies ist insofern positiv zu werten, als dass eine stark zeitversetzte Rückmeldung dazu führen könnte, dass ein Testverfahren zum Zeitpunkt der Rückmeldung für Lehrkräfte bereits als abgeschlossen gilt und die Testergebnisse aus diesem Grund gar nicht mehr zur Kenntnis genommen werden (Skejic et al., 2015). Ein weiterer Unterschied besteht in der Auswahl der Stichprobe, da es sich bei VERA, im Gegensatz zu bspw. PISA oder dem IQB-Bildungstrend, die auf einer repräsentativen Stichprobe basieren, um eine Vollerhebung handelt (Sälzer, 2016).

Bereits in ihrer Zielsetzung unterscheiden sich die Vergleichsarbeiten von den genannten Instrumenten PISA und IQB-Bildungstrend. Während diese eindeutig als Instrumente des Bildungsmonitorings auf internationaler oder nationaler Ebene zu werten sind, entfalten die Vergleichsarbeiten ihre Wirkung auf Ebene der Schule, mit dem Ziel der Schul- und Unterrichtsentwicklung (Maaz et al., 2019; Sälzer, 2016). Internationale Vergleichsstudien auf Systemebene wie PISA sind dagegen wenig geeignet, Aussagen über einzelne Schulen zu treffen oder gar die Individualebene der einzelnen Schüler*innen zu beleuchten (Gathen, 2006). Mit dem Ziel, Leistungswerte für ganze Schulsysteme bzw. Subsysteme (z. B. Schulformen) für internationale Ländervergleiche zu erhalten, erfolgt die Testentwicklung und Berichtslegung derartiger System-Monitoring-Studien basierend auf einem entsprechenden Testmodell (Bonsen et al., 2006). Der IQB-Bildungstrend basiert zwar ebenso wie VERA auf Kompetenzstufenmodellen für die Jahrgangsstufe 4 des Primarbereichs und Jahrgangsstufe 9 der Sekundarstufe I, Ziel ist jedoch eine langfristige Trendbeobachtung bzw. Vergleiche zwischen den Bundesländern im Sinne eines Bildungsmonitorings, ähnlich wie bei PISA, lediglich auf Bundesebene (KMK, 2016; Maaz et al., 2019; Sälzer, 2016).

2.2.1.6. Low-/No-Stakes vs. High-Stakes

Vergleichsarbeiten unterscheiden sich von vielen ähnlich gestalteten Tests im internationalen Kontext auch insofern, dass es sich bei den Vergleichsarbeiten um *low-stakes* Tests handelt, also Tests ohne Konsequenzen für die beteiligten Akteur*innen (Ramsteck & Maier, 2015). Generell zeichnen sich die im deutschen Schulsystem etablierten Systeme zur Rechenschaftslegung (*accountability regimes*), zu denen auch die Vergleichsarbeiten zählen, durch ihren no- bzw. low-stakes Charakter aus (C. Thiel, Schweizer & Bellmann, 2017). Bei derartigen Tests gilt das durch die Tests gewonnene Informationsangebot als ausreichender Anreiz für schulische Akteur*innen, sich mit den Testergebnissen auseinanderzusetzen und entsprechende Maßnahmen abzuleiten (Altrichter et al., 2016).

Im US-amerikanischen Schulsystem etablierte Tests zur Überprüfung schulischer Mindeststandards fallen dagegen i. d. R. in die Kategorie der *high-stakes* Tests. In diesen Systemen werden standardisierte Tests zur Leistungsbeurteilung von Lehrkräften und Schulen genutzt, häufig in Verbindung mit disziplinarischen Konsequenzen und schulischen Sanktionen, wie dem Entzug finanzieller Mittel, Strafversetzungen oder gar Entlassungen bei wiederholt unterdurchschnittlichen Leistungen der Schüler*innen (Ramsteck & Maier, 2015). Der dadurch entstehende *Accountability*-Druck, zusätzlich befeuert durch die Veröffentlichung der Testdaten, soll eine

Qualitätsentwicklung anregen. Bei den Vergleichsarbeiten – als low-stakes Instrument – wird *Accountability* hingegen als professionelle Selbstregulierung und - kontrolle realisiert. Es stehen hierbei keine Sanktionsmaßnahmen im Raum (Bach, Wurster, Thillmann, Pant & Thiel, 2014).

Im Forschungskontext der high-stakes Testungen existiert im internationalen Bereich ein umfassender Korpus an Forschungsarbeiten, die häufig auch negative und nicht erwünschte Effekte derartiger Testsysteme aufzeigen (Maier & Kuper, 2012). Die Handlungslogik dieser *high-stakes* Testsysteme folgt der Prämisse, dass Lehrkräfte und Schulen durch die Überprüfung des Erreichens festgesetzter Leistungsstandards, verknüpft mit Belohnungen und Sanktionen, motiviert werden und Leistungsrückmeldungen für eine datengestützte Unterrichtsentwicklung nutzen (Maier, 2007). In der Praxis scheint diese Rechnung jedoch nicht immer aufzugehen, werden doch in der angloamerikanischen Literatur häufig auch negative Konsequenzen externer Evaluationsverfahren mit high-stakes Charakter zitiert (siehe bspw. Amrein & Berliner, 2002; Berliner, 2011; Gulek, 2003; B. D. Jones & Egley, 2004). Als solche nicht intendierte Effekte werden die gezielte Vorbereitung von Schüler*innen durch Lehrkräfte (*Teaching to the Test*), eine Verengung der Unterrichtsinhalte auf testrelevante Themen (*Narrowing the Curriculum*) oder opportunistisches Verhalten von Lehrkräften, bspw. durch Hilfestellungen während der Tests, hervorgehoben (Demski, 2019b; Maier, 2007).

Zwar liegen im englischsprachigen Raum gemäß der weiter zurückreichenden Testtradition entsprechend mehr Untersuchungen zu externen Leistungsprüfungen vor, jedoch lassen sich die dort gewonnenen Erkenntnisse zum high-stakes Testing nicht einfach auf das deutsche Schulsystem und speziell die hier implementierten Vergleichsarbeiten übertragen (Dedering, 2011; Maier, 2010b).

Da der Fokus dieser Arbeit explizit auf dem Testinstrument VERA als charakteristisches low-stakes Instrument liegt, sei an dieser Stelle bei Interesse an weiteren Informationen zum high-stakes Testing exemplarisch auf folgende Arbeiten verwiesen, die sich u. a. mit den intendierten und nicht-intendierten Effekten von high-stakes Testing auseinandersetzen: Maier (2010b) spiegelt in einem systematischen Literaturreview den internationalen Forschungsstand zur Wirkung testbasierter Rechenschaftslegung wider. Die Arbeiten von Cizek (2001), G. M. Jones, Jones und Hargrove (2003) und Nichols und Berliner (2005) beschäftigen sich ausführlich mit den unbeabsichtigten Konsequenzen von high-stakes Testing.

Auch unter low-stakes bzw. no-stakes Bedingungen ist jedoch u. U. mit unbeabsichtigten Nebenfolgen zu rechnen (Bellmann, Schweizer & Thiel, 2016). Bellmann und Weiß (2009) sprechen von zahlreichen anderen möglichen Effekten, die bis zu Betrug durch die Lehrkraft reichen. Demnach zeigen sich bei der Durchführung von Vergleichsarbeiten nicht nur – mehr oder weniger ausgeprägt – die damit intendierten Effekte in Form von Maßnahmen zur Schul- und Unterrichtsentwicklung, sondern u. U. auch nicht intendierte Effekte, respektive Maßnahmen, die nicht auf Qualitätsentwicklung, sondern ausschließlich auf ein möglichst gutes Testergebnis abzielen. Jedoch scheinen sich die Befürchtungen derartiger eher negativ zu bewertender Auswirkungen aufgrund der durch den low-stakes Charakter bedingten Absenz negativer Konsequenzen und Sanktionen, sowohl für Schüler*innen als auch für Fachlehrkräfte, nur in Einzelfällen zu bewahrheiten (Richter, 2016). Empirische Nachweise für starke testbedingte nachteilige Unterrichtsveränderungen finden sich demnach nicht (Maier, 2010b).

2.2.1.7. Erklärungsmodelle zur Datennutzung

Wie immer wieder in der Literatur herausgestellt, ist für eine durch VERA angestoßene Schul- und Unterrichtsentwicklung eine produktive Nutzung der rückgemeldeten Daten unerlässlich (Diemer & Kuper, 2011; Hartung-Beck & Diemer, 2009). Die Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung beschreibt dahingehend den idealtypischen Kreislauf datengestützter Unterrichtsentwicklung: Nach der Testdurchführung erfolgt demnach im Idealfall eine systematische Auswertung in den Fachgruppen, gefolgt von einem Austausch der Lehrkräfte über mögliche Ursachen der festgestellten Ergebnisse und einer daran angeschlossenen gemeinsamen Ausarbeitung und Festlegung von Zielen und Maßnahmen. Der Erfolg dieser im Anschluss an die Vergleichsarbeiten von den jeweiligen Lehrkräften umgesetzten Maßnahmen wird dann im letzten Schritt mit der erneuten Testung evaluiert (KMK, 2010).

Ein derartiges Kreislaufschema wird u. a. im *Rahmenmodell zur pädagogischen Nutzung von Vergleichsarbeiten* von Helmke (2004) bzw. Helmke und Hosenfeld (2005) aufgegriffen. Dieses zählt neben dem Modell wichtiger Einflussfaktoren von *School Performance Feedback Systems (SPFS)* von Visscher und Coe (2002, 2003) zu den prominentesten Modellen zur Beschreibung der pädagogischen Nutzung empirischer Leistungsdaten von Schüler*innen. Beide Modelle werden in Forschungsarbeiten im VERA-Kontext häufig herangezogen, setzen dabei jedoch einen unterschiedlichen Fokus. Während das SPFS-Modell von Visscher und Coe (2002, 2003) verschiedene Einflussfaktoren auf die schulinterne Nutzung eines School Performance

Feedback Systems aus Sicht der Einzelschule beschreibt, skizziert das Modell von Helmke und Hosenfeld (2005) den idealtypischen Prozess der Evaluationsdatennutzung durch die individuelle Lehrperson.

Da das Forschungsinteresse dieser Arbeit auf der Perspektive der Lehrkräfte liegt und auch für das Verständnis des in Kapitel 2.2.2 folgenden Forschungsstandes relevant ist, wird hier nur das Rahmenmodell zur pädagogischen Nutzung von Vergleichsarbeiten nach Helmke und Hosenfeld (2005) vorgestellt (siehe Abbildung 6).

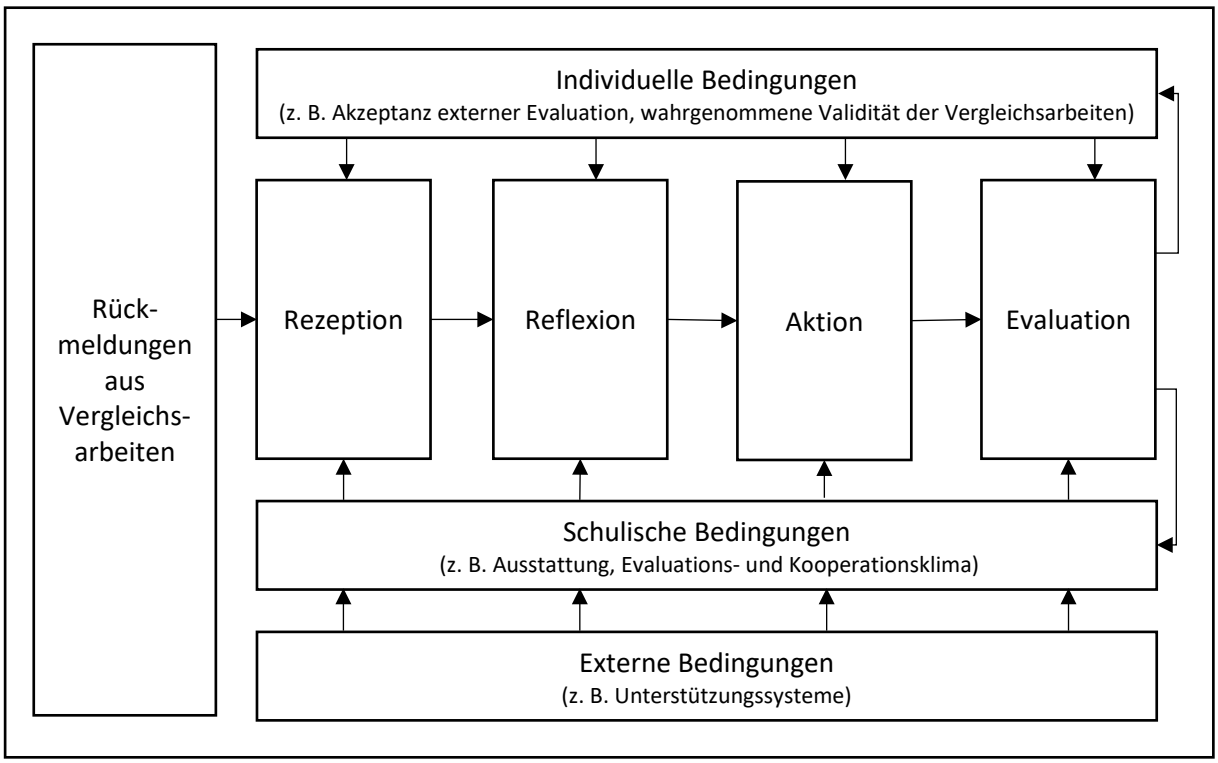


Abbildung 6: Prozessmodell zur pädagogischen Nutzung von Vergleichsarbeiten nach Helmke (2004) und Helmke und Hosenfeld (2005). Vereinfachte Darstellung in Anlehnung an Koch (2011)

Ausgangpunkt des Modells bilden durch das externe Evaluationsverfahren generierte und in den Rückmeldungen der Vergleichsarbeiten enthaltene Informationen über Leistungen und Kompetenzen der Schüler*innen, einschließlich verschiedener Vergleichsinformationen bzw. Leistungsreferenzen. In einem ersten idealtypischen Verarbeitungsschritt erfolgt die Rezeption der Rückmeldungen, indem sich die Lehrkraft mit deren Inhalten auseinandersetzt und daraus ein Verständnis generiert. Daraufhin erfolgt eine detaillierte Analyse der Daten (Reflexion), aus welcher sich wiederum Maßnahmen der Unterrichtsveränderung ableiten (Aktion). Die Wirksamkeit der ergriffenen Maßnahmen wird sodann durch eine erneute Evaluation überprüft. Das

Modell beschreibt somit auf der horizontalen Achse einen schrittweisen Prozess bzw. Zyklus aus Rezeption, Reflexion und Aktion, gefolgt von einer wiederholten Evaluation. Zusätzlich werden alle Phasen von verschiedenen Kontextfaktoren mitbeeinflusst. Das Modell benennt hier zum einen individuelle Bedingungen, wie bspw. Akzeptanz des Evaluationsverfahrens oder Vorwissen und Expertise der Lehrkraft, zum anderen schulische Faktoren, wie die Kooperation im Kollegium, und auch externe Faktoren, wie die Bereitstellung von Unterstützungssystemen (Helmke, 2004; Helmke & Hosenfeld, 2005).

In dieser differenzierten Darstellung individueller, schulischer und externer Bedingungen, die den Verarbeitungsprozess der Rückmeldeinformationen beeinflussen, sehen Altrichter et al. (2016) die Stärke dieses Prozessmodells. Allerdings gibt es auch kritische Stimmen, besonders in Bezug auf die theoretische Fundierung des Modells (bspw. Maier, 2009d). Demnach mangelt es an einer ausreichenden inhaltlichen Spezifikation von Zusammenhängen, also einer Konkretisierung inhaltlicher Kausalzusammenhänge zwischen den verschiedenen Einflussfaktoren und der Nutzung der Daten, was einer empirischen Modellprüfung im Wege steht (Demski, 2016; Maier, 2008a). Zumindest Teilaspekte des Modells werden jedoch in den Arbeiten von Koch (2011) und Groß Ophoff (2013) untersucht und empirisch untermauert.

2.2.1.8. Der Kreislauf der Datennutzung in der Praxis

Trotz verschiedenster Kritikpunkte verdeutlicht das Prozessmodell von Helmke und Hosenfeld sehr gut, dass Leistungsrückmeldungen bzw. -informationen nicht per se zu Veränderungen und Entwicklungen führen, sondern dass eine datenbasierte Unterrichtsentwicklung durchaus voraussetzungsvoll ist (Altrichter et al., 2016; Zimmer-Müller et al., 2014) und ein, wie von der KMK (2010) formulierter Kreislauf zur datenbasierten Unterrichtsentwicklung von verschiedenen Faktoren abhängig ist (Wurster & Richter, 2016). So folgt aus der Rückmeldung der Daten nicht zwingend deren schulinterne Nutzung. Eine produktive Nutzung der erhaltenen empirischen Informationen erfordert eine Transformation in Handlungswissen durch Lern-, Entscheidungs- und Handlungsprozesse der betroffenen schulischen Akteur*innen (Altrichter et al., 2016). Auch Zuber (2019) betont, dass die erfolgreiche Wirkungsentfaltung jeglicher Elemente der Neuen Steuerung, also auch der Vergleichsarbeiten, die Partizipation der involvierten Akteur*innen verlangt. Mit Blick auf die Forschungsliteratur der vergangenen Jahre zieht die Autorin jedoch ein eher ernüchterndes Fazit: Die Beteiligung bleibt bisher deutlich hinter den teils ambitionierten Erwartungen zurück. Lehrkräfte orientieren sich zwar durchaus

in ihrer Unterrichtsgestaltung am Bildungsstandard, eine Auseinandersetzung mit oder die Verwendung von Rückmeldedaten aus standardisierten Leistungstests findet dagegen kaum statt.

Die Lücke zwischen Theorie und Praxis beginnt schon bei der Wahrnehmung der Funktion von Vergleichsarbeiten. Auch wenn die Ziele von VERA formal klar definiert sind, zeigen Forschung und Praxis, dass die Wahrnehmung von Sinn und Zweck der Vergleichsarbeiten auf Seiten der betroffenen schulischen Akteur*innen von dieser Definition abweicht. VERA wird zwar als Instrument zur externen Leistungsdiagnose wahrgenommen, allerdings kaum als hilfreich für eine Unterrichtsverbesserung bzw. -entwicklung empfunden (Richter, 2016). Insgesamt wird die Nützlichkeit der Vergleichsarbeiten in der pädagogischen Praxis häufig in Frage gestellt (Graf et al., 2016) (siehe auch Kapitel 2.2.2.3 zur wahrgenommenen Nützlichkeit).

Die Nutzung der durch die Testung gewonnen empirischen Daten variiert teils stark, und das Instrument stößt häufig auf Ablehnung. Auf Ebene der Schulen werden die vorliegenden Daten kaum genutzt, um Schulentwicklungsprozesse anzustoßen (Ramsteck, Muslic, Graf, Maier & Kuper, 2015; Ramsteck & Maier, 2015). Hinzu kommt, dass viele Lehrkräfte bereits die Durchführung und Auswertung der Vergleichsarbeiten als zusätzliche Belastung neben dem schulischen Alltagsgeschäft empfinden, obwohl mit der Rückmeldung der Ergebnisse erst die eigentliche Arbeit im Sinne einer Förderung der Unterrichtsentwicklung und der Förderung schulischer Entwicklungsprozesse beginnt (Skejic et al., 2015). Eine evidenzbasierte Schulentwicklung scheint speziell im deutschen Sprachraum eine besondere Herausforderung für Lehrkräfte darzustellen (Gathen, 2006).

An diese Ausgangslage schließt sich nun die Frage nach den Gelingensbedingungen pädagogischer Instrumente – wie den Vergleichsarbeiten – an. Unter welchen Umständen können die relevanten Akteur*innen für die Nutzung solcher Instrumente gewonnen werden? Daher wird im folgenden Kapitel 2.2.2 der Forschungsstand zur Perspektive der schulischen Akteur*innen auf die Vergleichsarbeiten dargestellt. Ein besonderes Augenmerk liegt hierbei auf der Sichtweise der Lehrkräfte und der speziellen Frage nach deren Akzeptanz von Vergleichsarbeiten.

2.2.2. Vergleichsarbeiten aus der Perspektive von Lehrkräften – Forschungsstand nach zwei Dekaden VERA-Forschung

Während die ersten Veröffentlichungen zu Vergleichsarbeiten zu Beginn der 2000er-Jahre noch vor allem beschreibenden und konzeptionellen Charakter haben und bspw. Van Ackeren und

Bellenberg im Jahr 2004 noch den dringenden Bedarf an begleitenden empirischen Forschungsarbeiten im Bereich der gerade in der Implementation befindlichen Vergleichsarbeiten anmahnen, entsteht in den darauffolgenden Jahren bzw. Jahrzehnten seit Einführung der Vergleichsarbeiten ein ganzer Forschungsstrang zur Untersuchung der Akzeptanz, Rezeption und Nutzung der rückgemeldeten Daten durch schulische Akteur*innen (Maier & Kuper, 2012).

In der aktuellen Forschungsliteratur lassen sich zur Perspektive der Lehrkräfte im Hinblick auf Vergleichsarbeiten folgende Kategorien bzw. Faktoren identifizieren, die nachstehend systematisch vorgestellt werden und anhand derer dieses Kapitel strukturiert wird: Akzeptanz, Einstellung, Nützlichkeit, zeitliche Belastung bzw. Aufwand-Nutzen-Abwägung, Nutzung der Ergebnisse und weitere Wahrnehmungsfaktoren. Hierbei sei angemerkt, dass sich die dargestellten Wahrnehmungsaspekte meist nicht klar trennen lassen und auch teils wechselseitige Beziehungen zwischen den einzelnen Bewertungsaspekten bestehen, die auch den Gegenstand der empirischen Untersuchung darstellen.

2.2.2.1. Akzeptanz

Definition von Akzeptanz und deren Bedeutung in der Forschung zu Vergleichsarbeiten

Der Begriff Akzeptanz wird im Kontext von VERA häufig gebraucht und meist als wichtige Voraussetzung einer auf den Vergleichsarbeiten basierenden erfolgreichen Schul- und Unterrichtsentwicklung angesehen (siehe bspw. Groß Ophoff et al., 2019; Kühle & Peek, 2007; Maier, 2008a, 2009c; 2010a; Vogel et al., 2016; Vogel, 2020; Wagner et al., 2019). Hier ist jedoch auffällig, dass i. d. R. in keiner der entsprechenden Arbeiten eine Definition des Akzeptanzbegriffs erfolgt. Auch das Prozessmodell der pädagogischen Nutzung von Vergleichsarbeiten von Helmke (2004) bzw. Helmke und Hosenfeld (2005) konzeptualisiert Akzeptanz als einen individuellen Bedingungsfaktor der Auseinandersetzung mit den Ergebnissen und der darauf aufbauenden Umsetzung von Verbesserungsmaßnahmen. Das Konstrukt Akzeptanz wird somit nur über seine Wirkungsweise auf andere Faktoren definiert, eine konkrete Begriffsklärung und Definition erfolgt nicht. In diesem Kontext beschäftigen sich u. a. die Arbeiten von Koch (2011) und Groß Ophoff (2013) mit der empirischen Aufarbeitung und zumindest partiellen Überprüfung des Prozessmodells von Helmke und Hosenfeld (2005) und beleuchten dabei auch den Einfluss der Akzeptanz als individuellen Wirkungsfaktor auf den Prozess der Rückmeldungsnutzung. Eine Konzeptualisierung oder Definition des Konstrukts Akzeptanz ist auch in diesen Arbeiten nicht auszumachen.

Die inhaltliche Bedeutung des Konstrukts lässt sich, neben seiner postulierten Wirkungsweise bzw. -richtung, nur über dessen Operationalisierung erschließen. Koch (2011) und Groß Ophoff (2013) greifen, ebenso wie verschiedene weitere quantitative Untersuchungen (siehe u. a. Groß Ophoff et al., 2019; Maier & Rauin, 2006; Maier, 2008a, 2008b, 2009d; 2010a; Maier, Bohl, Kleinknecht & Metz, 2011; Wagner et al., 2019; Wagner & Koch, 2021), auf eine von Ditton, Merz und Edelhäußer (2002) entwickelte Skala zur Messung von Einstellungen schulischer Akteur*innen gegenüber zentralen Tests zurück bzw. ergänzen diese teilweise um weitere Items. Ditton et al. (2002) sprechen jedoch explizit von Einstellungen, der Begriff Akzeptanz findet in dem entsprechenden Beitrag keine Verwendung. Die Operationalisierung von Akzeptanz mittels dieser Einstellungsskala liefert erste Hinweise darauf, dass Akzeptanz, analog zu den in Kapitel 2.1.2 aufgearbeiteten Einstellungstheorien wie speziell dem TAM, auch im Kontext der Vergleichsarbeiten eng mit Einstellungen verknüpft ist. Wagner et al. (2019) konstatieren konkret das Fehlen einer genauen Definition und verstehen Akzeptanz, in Anlehnung an das TAM, als eine positive Einstellung – im Falle dieser Untersuchung gegenüber Vergleichsarbeiten und Schulinspektion – und operationalisieren diese u. a. im Rückgriff auf Items der Einstellungsskala von Ditton et al. (2002) anhand der wahrgenommenen Nützlichkeit der betrachteten Verfahren. Eine weitere Konkretisierung des Begriffs erfolgt zwar nicht, jedoch wird hierbei auch die Bedeutung der Nutzenwahrnehmung als weiterer Wahrnehmungsfaktor deutlich, welche, gerade im Kontext von Vergleichsarbeiten, eng mit Akzeptanz und Einstellung verbunden ist und weitestgehend im nächsten Abschnitt behandelt wird. Zunächst liegt der Fokus jedoch auf Forschungsarbeiten zur Untersuchung der Akzeptanz.

Zwar lässt sich in der einschlägigen Forschungsliteratur zu Vergleichsarbeiten keine Definition von Akzeptanz ausmachen, jedoch kann bspw. aus den Arbeiten, welche die – teils auch abgewandelte – Skala von Ditton et al. (2002) nutzen, durchaus ein Begriffsverständnis abgeleitet werden. Items wie „Die Vergleichsarbeit ist für die Arbeit der Schulen sehr wichtig“, „... sollte regelmäßig durchgeführt werden“ oder „... ist eine wichtige Grundlage, um Unterricht weiterentwickeln zu können“, bspw. eingesetzt von Maier (2009d), implizieren ein Verständnis von Akzeptanz im Sinne einer positiven Haltung einschließlich einer positiven Nutzenwahrnehmung. Maier (2009d, S. 341), in dessen Arbeit zwar ebenfalls keine explizite Definition zu finden ist, umschreibt die allgemeine Akzeptanz mit „... dem generellen Nutzen dieses Instrumentariums für die Schul- und Unterrichtsentwicklung ...“ und setzt dadurch Akzeptanz in gewisser Weise mit Nützlichkeit gleich. Auch in weiteren Beiträgen des Autors erfolgt eine Vermischung der Begriffe Einstellung, Akzeptanz und Nutzen. Maier (2008b) bspw. benennt

zunächst die Einschätzung des allgemeinen Nutzens von Vergleichsarbeiten sowie negative Effekte und die Bewertung der curricularen Validität als Indikatoren der Akzeptanz, spricht dann bei der Operationalisierung wiederum von Einstellungen gegenüber den Vergleichsarbeiten. Die Messung dieser Einstellungen erfolgt mittels dreier Skalen: Neben der allgemeinen Akzeptanz nach der Skala von Ditton et al. (2002), welche zuvor als allgemeiner Nutzen erläutert wurde, werden die Wahrnehmung von Vergleichsarbeiten als Belastung sowie deren Lehrplanvalidität erfasst. Einerseits wird somit die Nützlichkeit neben anderen Aspekten als Indikator der Akzeptanz betrachtet, andererseits werden Einstellungen mithilfe der Akzeptanzskala nach Ditton et al. (2002) erhoben, welche in anderen Arbeiten (bspw. bei Wagner & Koch, 2021) und auch bei Maier (2008b) als Nützlichkeit interpretiert werden. Maier (2008a) nutzt dieselbe Operationalisierung, zunächst von Einstellungen gegenüber Vergleichsarbeiten, durch die drei zuvor genannten Skalen. Bei der Darstellung der Ergebnisse werden jedoch diese zuvor unter Einstellungen subsummierten Skalen unter der Kategorie Akzeptanz dargestellt, wodurch eine Vermischung der Begrifflichkeiten stattfindet. In einer weiteren Arbeit auf Grundlage derselben Konstrukte wird ein Verständnis von Akzeptanz als ein positiver Aspekt von Einstellung, neben curricularer Validität, als Antonym des Belastungsempfindens durch Vergleichsarbeiten deutlich (Maier, 2009c), eine Konzeptualisierung, die auch in der Arbeit von Vogel (2020) aufgegriffen wird.

Eine enge inhaltliche Verknüpfung zwischen Akzeptanz und Nützlichkeit zeigt sich auch in anderen Arbeiten. Bonsen et al. (2006) identifizieren zunächst Akzeptanz als eine wichtige Gelingensbedingung von Lernstandserhebungen, setzen jedoch, durch eine fehlende Erläuterung implizit ein Begriffsverständnis voraus bzw. sehen Akzeptanz in enger Verbindung mit Nützlichkeit. Daraus geht jedoch nicht hervor, ob Akzeptanz und Nützlichkeit gleichgesetzt werden oder Nutzenwahrnehmung eher als Voraussetzung bzw. Bedingungsfaktor von Akzeptanz gesehen wird. Wurster und Richter (2016) operationalisieren in einer Befragungsstudie mit Fachkonferenzleitungen Akzeptanz durch die wahrgenommene Nützlichkeit der untersuchten Verfahren, Vergleichsarbeiten und zentrale Abschlussprüfungen, vergleichbar mit Wurster und Bach et al. (2016), deren Untersuchungsfazit die konstatierte Nützlichkeitswahrnehmung der befragten Schul- und Fachkonferenzleitungen als Ausdruck von Akzeptanz interpretiert.

Wurster, Feldhoff und Gärtner (2016) dagegen verstehen Akzeptanz und Nutzenwahrnehmung als getrennte Konstrukte. Untersuchungsgegenstand sind jedoch nicht Vergleichsarbeiten, sondern Schulinspektionen. Akzeptanz erfasst dabei durch die Operationalisierung als allgemeine Diagnosegüte und die Bewertung der Ergebnisqualität für die individuelle eigene Schule eher

einen qualitativen Aspekt bzw. eine Bewertung der Tauglichkeit des betrachteten Instruments. Es geht bei diesem Akzeptanzverständnis somit eher um die Einschätzung der Fähigkeit eines Instrumentes, zu einem adäquaten und zutreffenden Urteil zu gelangen, als um die daraus resultierenden Konsequenzen. Diese werden durch Erhebung der Nutzenwahrnehmung und – im Fall dieser Studie – möglicher negativer Folgen abgedeckt. Auch die bereits benannten Arbeiten von Koch (2011) und Groß Ophoff (2013) differenzieren zwischen Nützlichkeit, bezogen auf die Bewertung der Rückmeldungen, und Akzeptanz, welche basierend auf der Interpretation der Items eher als die Wahrnehmung eines allgemeinen Nutzens bzw. Nützlichkeit angesehen werden kann, und operationalisieren diese als separate Konstrukte. Auch Groß Ophoff et al. (2019) begreifen Akzeptanz und Nützlichkeit als getrennte Konstrukte und erachten Akzeptanz als affektiv-kognitives Personenmerkmal.

Andere Ansätze einer Konzeptualisierung verfolgen bspw. Muslic (2017), die u. a. eine kritische Bewertung der Validität und Reliabilität des Testverfahrens als Indikator einer geringen Akzeptanz erachtet, oder Kühle und Peek (2007), die Akzeptanz als eine Art allgemeine resümierende Gesamtbewertung von Vergleichsarbeiten verstehen. Die genutzte Operationalisierung als Einzelitem „Abschließend betrachtet: Wie beurteilen Sie persönlich die Lernstandserhebungen im Allgemeinen?“ mit einer Skala von sehr positiv bis sehr negativ deckt die gesamte Breite von positiver bis negativer Bewertung ab, während andere Autor*innen, wie bspw. Wagner et al. (2019) oder Maier (2008a, 2009c), Akzeptanz als positives Konstrukt bzw. positiven Aspekt von Einstellung interpretieren.

Ein noch weitreichenderes Akzeptanzverständnis lässt sich in einem Projektbericht zur Evaluation der Nutzung von Bildungsstandardüberprüfung in den 8. Klassen österreichischer Schulen herauslesen. Anhand von im Rahmen des Evaluationsprojektes mit Schulleitungen geführten Interviews zeichnet sich – aus Sicht der Schulleitungen – im Hinblick auf die Akzeptanz ein allgemein positives Bild unter den Lehrkräften (Rieß & Zuber, 2014). Akzeptanz wird zwar auch hier nicht definiert, jedoch wird im Kontext von Akzeptanz von einer „... Bereitschaft zur Beteiligung an der Maßnahmenumsetzung ...“ (Rieß & Zuber, 2014, S. 37) gesprochen, was auf eine mögliche zugrundeliegende Definition von Akzeptanz als eine positive Haltung bis hin zu einer Handlungsbereitschaft schließen lässt.

Auch wenn der Begriff in der einschlägigen Literatur schwammig bleibt, finden sich zu der Frage, wie es um die Akzeptanz von Vergleichsarbeiten bei den schulischen Rezipient*innen,

speziell den Lehrkräften, bestellt ist, in der Forschungsliteratur zu Vergleichsarbeiten und Lernstandserhebungen zahlreiche Erkenntnisse, die in den folgenden Abschnitten erläutert werden.

Status Quo und Trends in der Akzeptanz der Lehrkräfte

In einer Survey-Studie mit baden-württembergischen Lehrkräften nach der ersten verpflichtenden Durchführung der Vergleichsarbeiten im Schuljahr 2005/2006 zeigt sich ein eher verhaltenes Bild der Akzeptanz der Lehrkräfte. Die Bewertung der Akzeptanz, gemessen mit Hilfe der drei Skalen Allgemeine Akzeptanz, Vergleichsarbeiten als Belastung und Lehrplanvalidität, fällt für alle drei Aspekte unterdurchschnittlich aus und liegt im Mittel zumindest leicht unter dem semantischen Median der Skala. Es zeigen sich signifikante Korrelationen der Akzeptanzskalen mit verschiedenen Nützlichkeitsindikatoren (Maier, 2008a), was wiederum die bereits dargelegte inhaltliche Nähe zwischen Nützlichkeit und Akzeptanz untermauert. Insbesondere die allgemeine Akzeptanz korreliert stark mit den verschiedenen Facetten des Nutzens. Darauf aufbauend findet Maier (2008b) in einer Befragung ebenfalls baden-württembergischer sowie thüringischer Lehrkräfte aus dem Jahr 2007 eine eher niedrige Akzeptanz bei baden-württembergischen Lehrkräften, während die allgemeine Akzeptanz in Thüringen sogar deutlich über dem semantischen Median liegt. Er liefert dabei auch Hinweise auf Landesunterschiede in der Wahrnehmung von Vergleichsarbeiten, eine Erkenntnis, die sich auch in der Arbeit von Maier (2010a) bestätigt. Maier und Rauin (2006) ermitteln bei Befragungen baden-württembergischer Sekundar- und Grundschullehrkräfte in den Jahren 2003-2006 bzw. 2005 eine tendenziell positive Haltung der Lehrkräfte gegenüber Vergleichsarbeiten. Die befragten Lehrkräfte weisen im Mittel positive Akzeptanzwerte auf, wobei Grundschullehrkräfte den Testungen aufgeschlossener gegenüberstehen als Sekundarlehrkräfte.

Bereits zwischen den Jahren 2005 und 2007 lässt sich mithilfe einer längsschnittlichen bzw. quasi-längsschnittlichen Analyse ein abnehmender Trend in der Akzeptanz ausmachen. Der Autor erachtet diesen negativen Trend als sehr kritisch, da gerade ein Instrument wie die Vergleichsarbeiten stark an das professionelle Selbstverständnis von Lehrkräften appelliert und eine aktive und freiwillige Weiterarbeit erfordert (Maier, 2009d). Ebenso ermittelt Koch (2011) bei einer Analyse der Daten aus Lehrkräftebefragungen zwischen 2005 und 2008 einen abnehmenden Trend der Akzeptanz, vergleichbar mit den Erkenntnissen von Groß Ophoff et al. (2019), die eine rückläufige Entwicklung der Akzeptanz gegenüber Lernstandserhebungen bestätigen. Datengrundlage auch dieser Untersuchung bilden jährlich, im Anschluss an die

Vergleichsarbeiten, über einen Zeitraum von zehn Jahren (2005-2015) durchgeführte Lehrkräftebefragungen.

Kuhn (2014) bemerkt, dass auch 10 Jahre nach Einführung der Vergleichsarbeiten ein nicht unerheblicher Anteil der Lehrkräfte dem Instrument kritisch gegenübersteht und spricht von einem Akzeptanzdefizit. Ein weniger negatives Bild der Akzeptanz hingegen zeichnet u. a. Muslic (2017) und gelangt auf Grundlage von Fallstudien zu der Erkenntnis, dass viele Lehrkräfte durchaus Akzeptanz gegenüber Vergleichsarbeiten besitzen, betont jedoch, dass diese dennoch in der schulischen und unterrichtlichen Arbeit vernachlässigt werden. Dies leitet zu der Frage der Wirkungsweise von Akzeptanz und den Folgen fehlender oder auch vorhandener Akzeptanz auf das Handeln schulischer Akteur*innen, insbesondere im Hinblick auf Entwicklungsaktivitäten in Bezug auf Unterricht und Schule.

Wirkung von Akzeptanz

Mit der Thematik der Wirkungsweise von Akzeptanz beschäftigt sich u. a. Maier (2009c) bei der Untersuchung der pädagogischen Relevanz verpflichtender Tests wie Lernstandserhebungen an Schulen. In der Analyse zeigen sich signifikant positive Zusammenhänge zwischen der allgemeinen Akzeptanz und der Einschätzung verschiedener Nützlichkeitsaspekte, wie der Bewertung der Relevanz der Leistungsrückmeldung, insbesondere im Hinblick auf diagnostische Aktivitäten, aber auch hinsichtlich professioneller Urteile und einer Verbesserung des Unterrichts. Signifikante Zusammenhänge werden auch für die anderen Skalen der Akzeptanz bzw. Einstellung, Vergleichsarbeiten als Belastung und curriculare Validität ermittelt, die ebenfalls mit den verschiedenen Nützlichkeitseinschätzungen korrelieren, wenn auch teils in geringerem Ausmaß.

In den Forschungsarbeiten zur Akzeptanz von VERA zeigen sich nicht nur Zusammenhänge zur wahrgenommenen Nützlichkeit, es zeichnet sich auch die Bedeutung von Akzeptanz als wichtige Voraussetzung für eine Nutzung der Ergebnisse für weiterführende Aktivitäten ab (Vogel et al., 2016). Kühle und Peek (2007) bspw. identifizieren mit Hilfe eines multivariaten Regressionsmodells Akzeptanz, neben wahrgenommener Nützlichkeit, als wichtigsten Prädiktor für eine Nutzungsbereitschaft der rückgemeldeten Ergebnisse. In Anlehnung an das Prozessmodell der Datennutzung von Helmke und Hosenfeld (2005) findet auch Koch (2011) signifikante Effekte der Akzeptanz auf die Rezeption der Ergebnisrückmeldungen, gemessen anhand der Verständlichkeit, sowie die Reflexion, gemessen anhand von Intensität der

Auseinandersetzung sowie Nützlichkeit. Auf Grundlage dieser Ergebnisse betont die Autorin die Bedeutung der Akzeptanz von Vergleichsarbeiten für einen darauf aufbauenden Innovationsprozess. Ebenfalls auf Grundlage des Zyklusmodells von Helmke und Hosenfeld (2005) ermitteln Groß Ophoff et al. (2019) Akzeptanz als positiven Prädiktor der Intensität der Ergebnisauseinandersetzung und, in einem noch größeren Ausmaß, der wahrgenommen Nützlichkeit. Der Beitrag basiert wiederum zu Teilen auf Erkenntnissen von Groß Ophoff (2013), die mit Hilfe eines Strukturgleichungsmodells Akzeptanz gegenüber flächendeckenden Schulleistungsstudien als positiven Prädiktor von Auseinandersetzung mit und Nutzung der Ergebnisrückmeldungen identifiziert. Diese werden anhand von Fortbildungsaktivität und Unterrichtsveränderung im Hinblick auf Planung und Gestaltung sowie Aufgabenkultur gemessen. Darüber hinaus erweist sich die Akzeptanz auch hier als bedeutender Einflussfaktor der wahrgenommenen Nützlichkeit. Zu einem analogen Schluss gelangen Wagner et al. (2019) bei einer Befragung niedersächsischer Lehrkräfte im Schuljahr 2015/2016 mit dem Ziel einer Untersuchung der Vorhersagefähigkeit von Akzeptanz und Ergebnisreflexion auf tatsächliche Unterrichtsveränderungen im Vergleich von Vergleichsarbeiten und Schulinspektion. Bei Vergleichsarbeiten stellt sich Akzeptanz dabei als stärkerer Prädiktor für Anschlusshandeln heraus als bei Schulinspektion und erklärt fast 15 Prozent der Varianz des Anschlusshandelns im Unterricht.

Einen anderen Aspekt der Wirkung von Akzeptanz beleuchten Wagner und Koch (2021). Während die meisten Arbeiten Akzeptanz als Voraussetzung für Folgeaktivitäten betrachten, beschäftigen sich die Autorinnen in einem der neueren Artikel zum Thema Vergleichsarbeiten, bei der Analyse von Befragungsdaten von Grundschullehrkräften der Jahre 2010 bis 2016, mit dem Einfluss von Akzeptanz und funktionaler Wahrnehmung auf testvorbereitende Maßnahmen von Vergleichsarbeiten. Die Autorinnen beschreiben zunächst die ambivalente Funktion von Testvorbereitung im Hinblick auf daraus resultierende intendierte und nicht intendierte Effekte und analysieren die Häufigkeit verschiedener Vorbereitungsstrategien. Die Untersuchung gelangt zu dem Schluss, dass sowohl wahrgenommene Funktion als auch Akzeptanz einen Einfluss auf testvorbereitende Maßnahmen haben. Insgesamt wirkt hierbei die Wahrnehmung von VERA als Kontrollinstrument oder als ein Instrument zum sozialen oder ipsativen Vergleich tendenziell negativ auf die Intensität testvorbereitender Maßnahmen, wohingegen Akzeptanz positive Effekte auf alle untersuchten Vorbereitungsmaßnahmen aufweist.

Der aktuelle Forschungsstand spricht folglich dafür, dass Akzeptanz sowohl vor der Testdurchführung, im Hinblick auf vorbereitende Maßnahmen, als auch im Nachgang der Testung, in

Bezug auf die Nutzung der Rückmeldungen, verhaltenswirksam ist. Darüber hinaus finden sich in der Literatur Hinweise auf Faktoren, die wiederum die Akzeptanz beeinflussen.

Einflussfaktoren auf die Akzeptanz

Aufgrund einer fortwährend herrschenden Skepsis von Lehrkräften gegenüber Vergleichsarbeiten beschäftigt sich Kuhn (2014) mit den Ursachen des Akzeptanzdefizits bei VERA. Der Autor erachtet u. a. die Ableitung der VERA-Aufgaben aus den Bildungsstandards der 4. und 10. Klasse als problematisch, da dies bei einer Testung in der 3. bzw. 8. Klasse insbesondere bei schwächeren Schüler*innen zu Überforderung und Demotivierung führt, was wiederum von den Lehrkräften kritisch beurteilt wird. Auch wirkt der Informationsgewinn aus VERA gerade in Brennpunktschulen für Lehrkräfte eher frustrierend. Bedingt durch Varianzen zwischen den Jahren, die für schulische Akteur*innen wenig plausibel erscheinen, ist es außerdem schwierig, die Ergebnisse für schulinterne Trendmessungen zu nutzen.

Einen weiteren nicht unerheblichen Punkt, der sich in verschiedenen Publikationen wiederfindet, bildet das in der Schulpraxis häufig als ungünstig wahrgenommene Verhältnis von Aufwand und Ertrag der Vergleichsarbeiten. Auch Diemer (2013) ermittelt in diesem Zusammenhang einen Zeitfaktor als Einflussgröße der Akzeptanz. Mit Hilfe qualitativer Analysen von Interviews werden bestimmte Faktoren identifiziert, die sich negativ auf die Akzeptanz und dadurch auch auf die Ergebnisnutzung zentraler Vergleichsarbeiten auswirken können. Dadurch wird eine Kausalkette zwischen Akzeptanz und Nutzung impliziert, ohne diese jedoch konkret zu benennen. Negative Auswirkungen auf die Akzeptanz können nach Aussagen der interviewten schulischen Akteur*innen aus bestimmten Umständen und Widrigkeiten entstehen, wie bspw. technische Komplikationen bei der Eingabe und Übertragung der Ergebnisse und dem damit verbundenen generellen Zeitaufwand, der zusammen mit anderen Faktoren zur Überlastung der Lehrkräfte führen kann (siehe auch Kapitel 2.2.2.5). Förderlich für die Akzeptanz der Lehrkräfte hingegen wirkt die Vermittlung eines Mehrwerts von Vergleichsarbeiten für die tägliche Arbeit der Lehrkräfte zur Qualitätssicherung und -entwicklung (Maaz et al., 2019).

2.2.2.2. Einstellung

Im Folgenden wird der Aspekt Einstellung, der auch schon in Verbindung mit Akzeptanz angerissen wurde, etwas ausführlicher aufgearbeitet. Ein sehr breit gefasstes Einstellungs-

verständnis findet sich bspw. bei Vogel et al. (2016) und Vogel (2020), die zur Messung der Einstellungen von Lehrkräften gegenüber Lernstandserhebungen die Skalen von Maier (2008a) nutzen. Neben allgemeiner Akzeptanz, Belastungsempfinden und Lehrplanvalidität werden jedoch auch Nutzenaspekte unter Einstellung subsumiert, die in der Arbeit von Maier (2008a) als getrennte Konstrukte betrachtet werden. Auch hier zeigt sich wieder die Vernetzung von Einstellung, Akzeptanz und Nützlichkeit. Eine Operationalisierung von Einstellungen als wahrgenommene Nützlichkeit findet sich auch bei Wurster und Bach et al. (2016). Diese stellt sich in deren Untersuchung wiederum als wichtige Voraussetzung für Anschlussaktivitäten heraus (siehe auch Kapitel 2.2.2.3).

Eine starke Vermischung der Begrifflichkeiten erfolgt auch in den Arbeiten von Maier (2008a, 2008b, 2009c, 2009d), wie bereits im Abschnitt zur Definition von Akzeptanz erläutert. Da die Konzeptualisierung von Einstellung in diesen Arbeiten mit den Konstrukten allgemeine Akzeptanz, Wahrnehmung von Vergleichsarbeiten als Belastung und Einschätzung der Lehrplanvalidität bereits in eben jenem Abschnitt dargestellt wurde, werden in diesem an dieser Stelle nur einige bisher noch nicht angesprochene Untersuchungsergebnisse dargestellt. Maier (2008a) bspw. untersucht u. a. den Zusammenhang der Einstellung von Lehrkräften und der Nutzenwahrnehmung von Vergleichsarbeiten. Neben der Erkenntnis einer eher verhaltenen Einschätzung von Einstellung bzw. Akzeptanz zeigen sich signifikante Zusammenhänge der einzelnen Konstrukte mit verschiedenen Facetten der Nutzeneinschätzung. Die Wahrnehmung der Vergleichsarbeiten als Belastung steht u. a. in einem negativen Zusammenhang mit der Nutzenwahrnehmung. VERA wird somit weniger als Belastung wahrgenommen, wenn Lehrkräfte einen Nutzen in der Durchführung sehen. Dieser Befund liefert weitere Hinweise für die enge Verknüpfung von Einstellung bzw. Akzeptanz und wahrgenommener Nützlichkeit. Darüber hinaus zeigen sich Bewertungsunterschiede nach Schulform und Fach, eine Frage, die auch Maier (2009c) beschäftigt. Im Hinblick auf die Fragestellung, wieweit die Einstellungen von Lehrkräften gegenüber verpflichtenden Tests vom Schultyp und dem Unterrichtsfach abhängig sind, zeigen sich in einer quantitativen Befragung von Lehrkräften weiterführender Schulen Unterschiede in der Wahrnehmung von und Einstellung zu Vergleichsarbeiten bei Lehrkräften verschiedener Schultypen. Lehrkräfte in Hauptschulen erweisen sich generell aufgeschlossener gegenüber verpflichtenden landesweiten Tests als ihre Kolleg*innen an Realschulen oder Gymnasien. In Abhängigkeit der Schulart zeigen sich zusätzlich Bewertungsunterschiede je nach getestetem Fach. Fächerunterschiede finden sich auch bei Maier (2009d), wobei Mathematiktests fast durchweg positiver bewertet werden als die Tests im Fach Deutsch,

eine Erkenntnis, die sich auch schon bei Maier (2008b) zeigt. Schulformeffekte treten in dieser Untersuchung dagegen nur bei einzelnen Konstrukten auf.

Ein zunächst eher allgemeines Verständnis von Einstellung legen Skejic et al. (2015) bei der Untersuchung der Haltung von Fremdsprachenlehrkräften gegenüber VERA8 zugrunde, in der sie u. a. die Einstellungen von Lehrkräften zu Lernstandserhebungen beleuchten. Einstellungen werden in dieser Arbeit jedoch nicht definiert, können aber als eine Art allgemeine Bewertung im Sinne einer zustimmenden oder ablehnenden Haltung gegenüber Vergleichsarbeiten verstanden werden. Die Operationalisierung umfasst Items wie „Die Lernstandserhebungen leisten einen Beitrag zur langfristigen Qualitätsentwicklung an meiner Schule“, „Ich halte die Lernstandserhebungen insgesamt für eine sinnvolle Maßnahme“ oder „Die Lernstandserhebungen sind ein nützliches Instrument zur Lernstandsdiagnose“, die auf einer 5-stufigen Skala von Doppel-Plus bis Doppel-Minus erhoben werden. Der Fokus der Items auf Nutzenaspekte von Vergleichsarbeiten lässt wiederum auf eine starke implizite inhaltliche Verknüpfung zwischen Einstellung und Nützlichkeit schließen, wie sie bereits im Abschnitt zur Nutzung des Akzeptanzbegriffs bei VERA (siehe Kapitel 2.2.2.1) aufgezeigt wurde. Bei der Datenerhebung mit hessischen Lehrkräften ergibt sich, dass rund 40 % der Befragten Vergleichsarbeiten und deren Nutzen für die schulische Arbeit eher kritisch sehen. Etwa die Hälfte der Befragten bezweifelt, dass von den Vergleichsarbeiten Impulse für pädagogische Diskussionen ausgehen und dass diese zur langfristigen Qualitätsentwicklung an Schulen beitragen. Etwa ein Drittel der Befragten nimmt eine eher neutrale Haltung ein, was die Autor*innen auf einen potenziellen weiteren Informationsbedarf dieser Lehrkräfte zurückführen. Ein weiteres Drittel nimmt den Vergleichsarbeiten gegenüber eine positive Position ein. Jeweils 40 % empfinden VERA als Möglichkeit, über den eigenen Tellerrand zu blicken, und sehen in den Lernstandserhebungen ein nützliches Instrument zur Lernstandsdiagnose. Die Autor*innen bemerken jedoch, dass die Lehrkräfte an diesem Punkt häufig nicht weiterzuarbeiten scheinen und basierend auf der Diagnose kaum Unterrichts- oder Schulentwicklungsprozesse angestoßen werden. Insgesamt zeichnet sich mit Blick auf die Einstellung der Lehrkräfte eine heterogene Bewertung ab, die sowohl positive als auch negative Einschätzungen umfassen kann, wobei auch aus einer positiven Einstellung bzw. Nutzenwahrnehmung nicht zwangsläufig auf daran anschließende Aktionen geschlossen werden kann.

In der Arbeit von Diemer (2013) zeigt sich auch eine Vermischung der Begrifflichkeiten von Einstellung und Akzeptanz. Unter Rückgriff auf das Nutzungsmodell von Helmke und Hosenfeld (2005) als theoretische Grundlage, analysiert der Autor u. a. individuelle Bedingungs-

faktoren der Ergebnisnutzung. Während das Prozessmodell explizit von Akzeptanz als einem der individuellen Einflussfaktoren auf den Nutzungsprozess spricht (Helmke & Hosenfeld, 2005), benennt Diemer (2013) Einstellungen, neben anderen, als individuelle Bedingungsfaktoren. Anhand der qualitativen Auswertung von rund 70 Interviews werden verschiedene Kategorien von Einstellungen identifiziert. Zunächst erfolgt eine Unterscheidung zwischen sach- und personenbezogenen Bewertungen, die jeweils sowohl positive als auch negative Aspekte beinhalten. Bringt eine Lehrkraft eine positive sachbezogene Bewertung zum Ausdruck, werden Lernstandserhebungen allgemein für gut befunden bzw. akzeptiert. Neben dieser sehr allgemeinen Bewertung zeigen sich auch spezifischere funktionsbezogene Einstellungen. Hierzu zählen bspw. die positive Beurteilung von durch VERA entstehende schulinterne und -externe Vergleichsmöglichkeiten sowie der zentrale und unabhängige Prüfungscharakter oder die Wahrnehmung der Ergebnisse als aussagekräftig und hilfreich für die Unterrichtsgestaltung. Negative sachbezogene Bewertungen beziehen sich meist darauf, dass Sinn und Nutzen von Lernstandserhebungen eher bezweifelt werden. Lehrkräfte treffen diesbezüglich Aussagen zu Form und Inhalt der Testaufgaben, wie die Kritik an Fehlern und Unklarheiten der Aufgabenstellung, und betonen eine vom Unterricht abweichende Schwerpunktsetzung ebenso wie mangelnde curriculare Validität und die zeitliche Belastung insbesondere durch Korrektur und Datenübermittlung.

Zu den Einstellungen im Sinne personenbezogener Bewertungen zählen ein positiver Affekt, wie z. B. Freude über das gute Abschneiden der eigenen Klasse oder die Einschätzung der Vergleichsarbeiten als persönliche Bereicherung sowie als hilfreich für die Reflektion der eigenen Arbeit. In negativen Einschätzungen bringen Lehrkräfte ihren Ärger über VERA zum Ausdruck, bspw. weil die Vergleichsarbeiten als nicht aussagefähig oder nutzlos eingeschätzt werden. Einige Lehrkräfte sprechen in diesem Zusammenhang auch von einer als überflüssig empfundenen zeitlichen Belastung oder Angst vor Schuldzuschreibungen bei schlechtem Abschneiden der eigenen Klasse. Nach der Interpretation des Autors beschreiben die beiden durch die Interviews identifizierten Kategorien mentale bzw. emotionale Zustände. Mit dieser Unterscheidung verschiedener Facetten von Einstellungen lassen sich Parallelen zu den in Kapitel 2.1.1 beschriebenen kognitiven und affektiven Komponenten der Einstellung ziehen, welche auch im Kontext der Untersuchung von Datennutzung bzw. Bildungsstandardreformen bei Vanhoof, Vanlommel, Thijs und Vanderlocht (2014) und Zuber (2019) zu finden sind. Zuber (2019) benennt hier bspw. Wut, Ärger, Misstrauen und Freude als affektive Ausprägungen von

Einstellungen, Aspekte, die in Kapitel 2.2.2.6 teils noch einmal ausführlicher aufgegriffen werden.

In diesem Zusammenhang betont Diemer (2013), dass eine teils negative Bewertung nicht zwangsläufig eine vollständige Ablehnung des Verfahrens bedeutet, sondern neben einer eher negativen Einschätzung bestimmter Aspekte gleichzeitig andere durchaus positiv bewertet werden können. Dies spricht für ein ambivalentes Verständnis von Einstellungen, wonach diese sowohl positive als auch negative Facetten in der Bewertung eines Einstellungsobjektes beinhalten können. Die Ambivalenz von Einstellungen spielt auch bei Zuber (2019) eine Rolle. Die Autorin ermittelt in einer ebenfalls qualitativen Untersuchung häufig zwiespältige Einstellungen von Lehrkräften gegenüber Bildungsstandardreformen im österreichischen Schulsystem. Die Ambivalenz rührt dabei von der unterschiedlichen Bewertung verschiedener Aspekte her, die jedoch, nach Einschätzung der Autorin, gerade in quantitativen Arbeiten i. d. R. keine Berücksichtigung finden, da Einstellungswerte dort meist gemittelt werden, weshalb eine Differenzierung einzelner Aspekte durch Lehrkräfte meist nicht sichtbar wird.

Obwohl sich Zubers (2019) Beitrag nicht direkt mit Vergleichsarbeiten beschäftigt, ist dieser dennoch für diese Arbeit interessant, weil die Autorin ihrer Untersuchung als eine der wenigen mit der Theory of Planned Behavior (TPB) eine Einstellungstheorie zugrunde legt und auf Grundlage derer das Verhältnis von Einstellung und Verhalten schulischer Akteur*innen untersucht, also eine Beziehung, die auch in dieser Arbeit beleuchtet wird (siehe Kapitel 2.1.2.3 für die ausführliche Beschreibung der TPB). Zuber (2019) untersucht unter Rückgriff auf die TPB die Wirkung von Einstellung, subjektiver Norm und wahrgenommener Verhaltenskontrolle auf das Umsetzungshandeln von Lehrkräften im Rahmen von Bildungsstandardreformen (siehe dazu auch Kapitel 2.2.2.6). Die Aufarbeitung des Forschungsstandes findet sowohl Belege für eine positive Beziehung zwischen Einstellung und Verhalten als auch Hinweise auf eine gänzliche Absenz eines Zusammenhangs. Anhand von Interviewdaten werden zunächst verschiedene Einstellungen ermittelt, wobei sich vielfältige, häufig auch ambivalente, kognitive und affektive Haltungen gegenüber Bildungsstandards zeigen. Interessant ist hier die Erkenntnis, dass positiv konnotierte Einstellungselemente stets kognitive Ausprägungen aufweisen, negative Äußerungen dagegen auch affektive Ausprägungen zeigen, wie bspw. die Befürchtung der Verfälschbarkeit und Kontrolle der Lehrtätigkeit, die auf ein Misstrauen auf Seiten der Lehrkräfte hinweist. Es zeigen sich keine rein positiven oder gar affektiv-positiven Einstellungen. Zusammenfassend findet sich, entgegen der theoretischen Annahmen der TPB, kein direkter Zusammenhang zwischen Einstellung und Handeln in Bezug auf Bildungsreformen auf der

Individualebene der Lehrkräfte. Auch bei negativen Einstellungen wurden Verhaltensintentionen geäußert, und umgekehrt.

Vanhoof et al. (2014) dagegen ermitteln auf Grundlage einer quantitativen Befragung flämischer Schulleitungen zur Datennutzung für Schulentwicklungsmaßnahmen Einstellungen als Prädiktoren von Datennutzung. In dem aufgestellten Pfadmodell wirken neben Selbstwirksamkeit sowohl affektive als auch kognitive Einstellungen signifikant auf die Datennutzung. Ebenfalls unter Bezugnahme auf die TPB stellt auch Demski (2019b) in einer Interviewstudie die Bedeutung der Einstellungen schulischer Akteur*innen als wichtigen Einflussfaktor evidenzbasierten Handelns heraus, wobei Einstellung wiederum implizit mit Nützlichkeit gleichgesetzt wird.

2.2.2.3. Nützlichkeit

Konzeption und Aspekte von Nützlichkeit

Die Analyse der Forschungsliteratur erlaubt den Schluss, dass eine ausreichende Nutzenwahrnehmung auf Seiten der schulischen Akteur*innen eine der wichtigsten Voraussetzungen der Nutzung von und Weiterarbeit mit Rückmeldungen aus Vergleichsarbeiten darstellt (siehe bspw. Bonsen et al., 2006; Kühle & Peek, 2007; Vogel, 2020; Wurster, Bach et al., 2016). In diesem Zusammenhang sei zu Beginn dieses Abschnitts darauf hingewiesen, dass in der Literatur zur Beschreibung desselben oder ähnlicher Sachverhalte sowohl von Nützlichkeit als auch von Nutzen die Rede ist. In dieser Arbeit werden die Begriffe daher synonym verwendet, wobei i. d. R. von Nützlichkeit gesprochen wird. Des Weiteren existieren auch im Hinblick auf die Nutzenwahrnehmung vielfach konzeptionelle Überschneidungen mit Einstellungen und Akzeptanz. Diese enge Beziehung zwischen den Konstrukten wird, wie bereits im Abschnitt zur Definition von Akzeptanz (siehe Kapitel 2.2.2.1) dargelegt, bei Wagner et al. (2019) deutlich, die die allgemeine Nützlichkeit als Teilaspekt von Akzeptanz betrachten. Die allgemeine Nützlichkeit erfasst demnach den wahrgenommenen Nutzen von Vergleichsarbeiten für Schul- und Unterrichtsentwicklung. Die Autor*innen weisen jedoch auf die Relevanz weiterer Nützlichkeitsaspekte hin, wie die individuelle Nützlichkeit der betrachteten Instrumente für die eigene pädagogische Arbeit. Diese Unterscheidung zwischen individueller, auf verschiedene Aspekte der eigenen Unterrichtsrealität bezogener und allgemeiner Nützlichkeit im Sinne eines Nutzens für die Schule bzw. den Unterricht im Allgemeinen findet sich mehrfach in der Literatur wieder (Dedering, 2011; Wurster, Feldhoff & Gärtner, 2016). Lehrkräfte empfinden hierbei den

individuellen Nutzen von Vergleichsarbeiten als Instrument zur Lernstandsdiagnose erkennbar größer als die allgemeine Nützlichkeit zur Unterrichts- bzw. Schulentwicklung (Skejic et al., 2015).

In der Forschungsliteratur gibt es zwei Zugänge zur Erfassung der Nützlichkeit von Vergleichsarbeiten. Zum einen wird die Einschätzung der Nützlichkeit des Verfahrens bzw. der gewonnenen Daten aus Sicht der schulischen Akteur*innen analysiert, sowohl auf allgemeiner Ebene als auch bezogen auf bestimmte schulische Bereiche. Zum anderen werden schulische Akteur*innen gebeten, verschiedene Nutzungsmöglichkeiten der Ergebnisse einzuschätzen (Dedering, 2011). Diese Unterscheidung wird jedoch meist nicht explizit getroffen, sondern ergibt sich implizit durch Operationalisierungen. Häufig werden die verschiedenen Ansätze durchmischt und zur Einschätzung des Nutzens sowohl Nützlichkeitsaspekte als auch die Bewertung von Nutzungsmöglichkeiten erhoben. Bonsen et al. (2006) bspw. erheben die Nützlichkeit verschiedener Aspekte der Ergebnisrückmeldung und analysieren wiederum deren Einfluss auf die Bedeutung der Ergebnisrückmeldungen für die Unterrichtsentwicklung. Die Nützlichkeit zeigt sich als signifikanter Prädiktor der Einschätzung der Bedeutung der Rückmeldungen für die Unterrichtsentwicklung.

Insgesamt nimmt die Untersuchung der Nützlichkeit von Vergleichsarbeiten häufig die Beurteilung diagnostischer Aspekte in den Blick. Nützlichkeit bezieht sich demnach darauf, dass Lehrkräfte durch die Vergleichsarbeiten zusätzliche diagnostische Informationen erhalten, die sonst derart nicht verfügbar wären, bspw. durch soziale Vergleiche mit Referenzgruppen und damit einhergehende erweiterte Möglichkeiten zur Leistungseinschätzung der eigenen Lerngruppe und zur Weiterentwicklung diagnostischer Kompetenzen (Spoden et al., 2014). Hier zeigen sich in den verschiedenen Forschungsarbeiten teils widersprüchliche, in jedem Fall jedoch stark heterogene Erkenntnisse. Die Aussagen einer Interviewstudie von Diemer (2013) lassen aus Sicht der befragten Lehrkräfte einen diagnostischen Mehrwert erkennen. Es kommt zum Ausdruck, dass die Rückmeldungen durchaus interessante Informationen liefern, die als Ergänzung sonstiger Informationen und Beurteilungen dienen oder vorherige Einschätzungen bestätigen. Diesbezüglich gibt es aber auch kritische Stimmen unter den Lehrkräften, die einen diagnostischen Nutzen von Vergleichsarbeiten in Frage stellen. Einige Lehrkräfte empfinden die durch VERA gewonnenen Informationen als überflüssig und redundant, wenn diese nur bisherige Leistungseinschätzungen bestätigen. Als aussagekräftig werden die Rückmeldungen insbesondere dann wahrgenommen, wenn diese einen zusätzlichen informativen Mehrwert haben.

Der Vorwurf der Redundanz wird auch in der qualitativen Arbeit von Jäger (2011) deutlich. Knapp über die Hälfte der 59 befragten Lehrkräfte gibt an, die eigenen Schüler*innen so gut zu kennen, dass keine externe Leistungsdiagnose notwendig und daher durch VERA keine neuen Erkenntnisse und Informationen, bspw. über Stärken und Schwächen, zu erhalten seien. Der andere Teil der Lehrkräfte betont dagegen durchaus, durch die Vergleichsarbeiten neue Anhaltspunkte zu erhalten. Dies bezieht sich vor allem auf einzelne Schüler*innen, deren Ergebnisse von den Erwartungen abweichen, und auf Hinweise zu bestimmten Inhaltsbereichen, die schon besonders intensiv oder noch gar nicht behandelt wurden. Ein Großteil der Befragten gibt an, dass ihrer Einschätzung nach in Folge der Vergleichsarbeiten keine Fördermaßnahmen möglich seien. Jeweils rund ein Drittel benennt mangelnde Zeit als Ursache für das Ausbleiben anschließender Fördermaßnahmen oder hält VERA generell für ein ungeeignetes Diagnoseinstrument, das lediglich bilanzierende Ergebnisdarstellungen ohne diagnostische Hinweise liefert. Auch in einer Untersuchung zur Bewertung und Nutzung von Ergebnisrückmeldungen im Rahmen österreichischer Bildungsstandards zeigt sich zwar durchaus eine Relevanz der Daten für diagnostische Zwecke, jedoch wenig Bedeutung für die Einleitung von Veränderungen, da viele Lehrkräfte nicht direkt wissen, wie sie aus den gewonnenen Informationen Erkenntnisse für ihren Unterricht ableiten können. Der Nutzen dieser zumindest mit VERA vergleichbaren Datenrückmeldungen zeigt sich somit eher im Bereich der Diagnose, weniger in der Handlungssteuerung (Grabensberger, Freudenthaler & Specht, 2008).

Eine sehr differenzierte Betrachtung von Nützlichkeit findet sich in vielen Arbeiten von Maier (2008a, 2008b, 2009c, 2009d, 2010a). Auf Basis eines umfassenden Itempools werden je nach Arbeit drei bis fünf Kategorien gebildet, die sich jedoch teils inhaltlich überschneiden. Der methodische Zugang zur Messung der Nützlichkeit aus Sicht der Lehrkräfte erfolgt über die Erfassung verschiedener Nutzungsmöglichkeiten von Vergleichsarbeiten. Unterschieden werden zunächst die Nützlichkeitsaspekte (förder-)diagnostische Nutzung, selektionsdiagnostische Nutzung und Hinweise für die Unterrichtsgestaltung (Maier, 2008a, 2009c). Die Kategorie diagnostische Nutzung beschreibt, inwieweit die Rückmeldungen von Vergleichsarbeiten als zusätzliche diagnostische Informationen und als nutzbare Unterstützung von Lerndiagnose, Beratung und Förderung wahrgenommen werden. Darunter fallen bspw. der Nutzen der Rückmeldungen zur Einschätzung von Stärken und Schwächen der Schüler*innen oder der Einsatz als Argumentationsgrundlage für Elterngespräche (Maier, 2008a, 2008b). Die zweite Kategorie zielt auf die Frage, ob Vergleichsarbeiten Lehrkräften zusätzliche Hinweise für die eigene Leistungsbewertung im Hinblick auf Notengebung oder Versetzungsentscheidungen liefern

können, bzw. ob Lehrkräfte ihre eigene Notengebung bestätigt sehen oder basierend auf den Testergebnissen hinterfragen (Maier, 2008b, 2009d). Maier (2010a) unterteilt diese Kategorie noch einmal explizit nach der Frage, ob Vergleichsarbeiten nützliche Hinweise für die eigenen Bewertungsmaßstäbe und -instrumente bieten (*Impact on professional judgement*) und der Frage nach dem Einfluss auf die Notengebung (*Impact on summative assessment*). Hinweise für die Unterrichtsgestaltung beziehen sich darauf, ob die Leistungsrückmeldungen aus Sicht der Lehrkräfte zu einer Reflexion über inhaltliche Schwerpunkte, bzw. eingesetzte Lehrmaterialien und Aufgabenstellungen führen (Maier, 2008a). Auch hier unterscheiden andere Arbeiten explizit die Subkategorien Hinweise für Unterrichtspraxis und Arbeitsaufträge und Hinweise auf curriculare Veränderungen (Maier, 2008b, 2009d, 2010a).

Weitere Aspekte der Nützlichkeit beleuchten Maier und Rauin (2006) und beziehen neben der Bewertung des pädagogischen Nutzens zusätzlich die Risiken von Vergleichsarbeiten als negative Aspekte in die Einschätzung der Nützlichkeit ein. Zu diesen Risiken zählen die Befürchtung des übermäßigen Übens für den Test, die Absenz eines Nutzens, da kein fairer Vergleich möglich ist, und die Einschätzung der eigenen Klassenarbeiten als ein besseres Messinstrument. Es stellt sich heraus, dass sowohl Grundschul- als auch Sekundarschullehrkräfte den förderdiagnostischen Nutzen von Vergleichsarbeiten überwiegend positiv einschätzen. Die Risiken werden jeweils von Lehrkräften ohne vorherige Testerfahrung kritischer gesehen. Insgesamt werden über beide Schularten die Risiken geringer bewertet als der Nutzen.

Insgesamt zeigt sich jedoch hinsichtlich der Nutzeneinschätzung der Lehrkräfte häufig ein eher verhaltenes Bild. In Maiers (2008a) Arbeit bspw. liegen die Skalenmittel aller drei Nützlichkeitsskalen unter dem semantischen Median von 3. Die höchste Zustimmung findet sich bei der selektionsdiagnostischen Nutzung, die geringste bei der förderdiagnostischen Nutzung. Der Autor führt diese geringe Nutzenwahrnehmung auf die Neuheit des Verfahrens zurück. Hinweise, die diese Annahme stützen, finden Groß Ophoff et al. (2019), die in der Einschätzung der wahrgenommenen Nützlichkeit, im Gegensatz zur Akzeptanz, zwischen den Jahren 2005 und 2015 eine tendenziell positive Entwicklung ausmachen und diese auf einen Gewöhnungseffekt der Lehrkräfte zurückführen.

Generell finden sich auch positive Bewertungen der Nützlichkeit. Kühle und Peek (2007) bspw. finden bei einer Lehrkräftebefragung im Schuljahr 2004/2005 eine 50-prozentige Zustimmung bei der Bewertung der Nützlichkeit. Im Jahr darauf geben sogar 64 Prozent der Lehrkräfte an, zumindest weitgehend vom Nutzen der Lernstandserhebungen überzeugt zu sein. Lehrkräfte

scheinen also durch die Rückmeldungen der Vergleichsarbeiten durchaus zusätzliche diagnostische Informationen zu erhalten. Jedoch variiert die Bewertung hinsichtlich der verschiedenen Aspekte. Am nützlichsten werden die rückgemeldeten Daten als Hinweise für zukünftig einzusetzende Lehrmaterialien und Aufgabenstellungen gesehen, Hinweise für curriculare Veränderungen ergeben sich in der Wahrnehmung der Lehrkräfte in den Anfangsjahren der Durchführung der Vergleichsarbeiten dagegen kaum.

Darüber hinaus scheint die Nutzeneinschätzung von Schulart und Bundesland abhängig zu sein (Maier, 2008b, 2009d, 2010a). Zusätzlich erweist sich die Einschätzung der Nützlichkeit als abhängig von der Unterrichtsnähe der befragten schulischen Akteur*innen: Schulleitungen bewerten die Bedeutsamkeit der Rückmeldungen positiver als Fachlehrkräfte (Dedering, 2011). Diese Erkenntnis teilen auch Bach et al. (2014), die aus ihrer vergleichenden Analyse schließen, dass Lehrkräfte die Nützlichkeit von Vergleichsarbeiten deutlich kritischer bewerten als Schulleitungen. Des Weiteren bestehen bei der wahrgenommenen Nützlichkeit Unterschiede hinsichtlich des unterrichteten Fachs. Dabei ist es auffällig, dass die Nützlichkeit in Mathematik insgesamt größer wahrgenommen wird als in den sprachlichen Fächern (Maier, 2009c; Vogel, 2020). Generell werden die Rückmeldungen von Vergleichsarbeiten eher dann als nützlich eingeschätzt, wenn eine Lehrkraft die getestete Klasse noch weiter unterrichtet und somit überhaupt eine Möglichkeit hat, mit den Ergebnissen weiterzuarbeiten (Jäger, 2011).

Wirkung von Nützlichkeit

Zwar unterscheiden sich die Einschätzungen schulischer Akteur*innen zur Nützlichkeit von Vergleichsarbeiten je nach Perspektive, Studiendesign, Fragestellung, Operationalisierung oder im Hinblick auf diverse Kovariaten, jedoch herrscht weitestgehender Konsens hinsichtlich eines Zusammenhangs von Nützlichkeit und Nutzung. In der Forschungsliteratur finden sich vielfach Hinweise für die Relevanz der Nutzenwahrnehmung der Vergleichsarbeiten für die Weiterarbeit mit den Rückmeldungen und eine anschließende Schul- und Unterrichtsentwicklung. Die wahrgenommene Nützlichkeit beeinflusst demnach nicht nur die Bereitschaft zu einer nutzungsorientierten Rezeption von Ergebnisrückmeldungen (Kühle & Peek, 2007), sondern erweist sich auch als signifikanter Prädiktor der tatsächlichen Ergebnisnutzung (Wurster & Richter, 2016). Auf Ebene der Schulleitungen entpuppt sich die Einschätzung der Nützlichkeit zudem als signifikanter Prädiktor einer Nutzung der VERA-Rückmeldungen für Personalentwicklungsmaßnahmen (Bach et al., 2014).

Koch (2011) folgt zur Untersuchung der Wirkung von Nützlichkeit wiederum der Logik des Prozessmodells von Helmke und Hosenfeld (2005) und fasst Nützlichkeit, neben Intensität der Auseinandersetzung, unter Reflexion der Ergebnisse. Es zeigt sich, dass diese beiden Konstrukte positiv miteinander korrelieren. Durch die Operationalisierung der Auseinandersetzungsintensität als ein Teil von Reflexion, neben Nützlichkeit, untersucht Koch (2011) zwar keinen Kausalzusammenhang zwischen den beiden Konstrukten, dennoch lassen sich aus den Erkenntnissen Hinweise auf einen Zusammenhang von Nützlichkeit und Ergebnisnutzung im Sinne einer Auseinandersetzung mit den Ergebnissen ableiten. Zusätzlich wirkt die wahrgenommene Nützlichkeit positiv auf verschiedene Aspekte der weiteren Ergebnisnutzung, hier unter Aktion gefasst. Somit wird Nützlichkeit, als qualitativer Aspekt von Reflexion, als wichtiger Schritt zur Erreichung von auf VERA basierenden Veränderungsprozessen identifiziert.

Eine zumindest ähnliche Operationalisierung wählt Groß Ophoff (2013) und ordnet die wahrgenommene Nützlichkeit, ebenso wie Intensität der Auseinandersetzung und Verständlichkeit, dem Konstrukt Auseinandersetzung mit den Rückmeldungen zu, welches wiederum die beiden Aspekte Rezeption und Reflexion umfasst. Die drei Konstrukte korrelieren signifikant positiv miteinander und liefern somit ebenfalls Hinweise auf einen Zusammenhang von Nützlichkeit und der Bereitschaft der Lehrkräfte, sich mit den Rückmeldungen auseinanderzusetzen. Einen positiven Zusammenhang zwischen wahrgenommener Nützlichkeit und Intensität der Auseinandersetzung bestätigen auch Groß Ophoff et al. (2019). Groß Ophoff (2013) findet darüber hinaus mittlere bis hohe Effekte für den Einfluss der wahrgenommenen Nützlichkeit auf Unterrichtsveränderungsaktivitäten und Fortbildungsaktivitäten in Folge der Vergleichsarbeiten.

Gemäß der Auffassung von Koch (2011) und Groß Ophoff (2013) ist Nützlichkeit somit bereits Teil des Auseinandersetzungsprozesses. Einen anderen Ansatz verfolgt bspw. Vogel (2020), nach dessen Verständnis Nützlichkeit einer Auseinandersetzung mit den Rückmeldungen und einer weiteren Nutzung vorgelagert ist. Der Autor unterscheidet die Nützlichkeitsperspektiven Lernstandsdiagnose, Unterrichtsentwicklung und Schulentwicklung und untersucht deren Prognosefähigkeit auf die Rückmeldenutzung. Es finden sich Hinweise auf die positive Wirkung der Nützlichkeit auf die Nutzung, gemessen anhand der Durchsicht der Rückmeldungen und der Intensität der Besprechung der Ergebnisse. Eine Schlüsselrolle der Nützlichkeitswahrnehmung bei der Nutzung datenbasierter Evaluationsrückmeldungen betonen auch Wurster und Bach et al. (2016). In einer multiplen Regressionsanalyse stellt sich die wahrgenommene Nützlichkeit bei allen schulischen Akteursgruppen, Schulleitungen, Fachkonferenzleitungen und Lehrkräften als bedeutendster Prädiktor der in Folge von VERA angestoßenen Entwicklungs-

aktivitäten heraus. Die Autor*innen identifizieren dabei einen dringenden Handlungsbedarf, wenn es um datenbasierte Weiterentwicklungen auf Schul- und Unterrichtsebene geht. Dass diese Forderung durchaus eine Berechtigung hat, bestätigen auch die Ergebnisse einer Befragung rheinland-pfälzischer Lehrkräfte und Schulleitungen, die verdeutlichen, dass die Nützlichkeit von Vergleichsarbeiten eher gering eingeschätzt wird und entsprechend selten aus Vergleichsarbeiten generierte Informationen für die schulische Weiterarbeit genutzt werden. Dies schlägt sich auch in einer starken Korrelation zwischen Nützlichkeitswahrnehmung und tatsächlicher Nutzung nieder, wobei sich die Nützlichkeit auch in dieser Untersuchung als wichtiger Prädiktor der Nutzung erweist (Demski, 2019b).

2.2.2.4. Nutzung der Ergebnisrückmeldungen

Wie bereits dargelegt, können Vergleichsarbeiten erst durch eine aktive Auseinandersetzung mit den rückgemeldeten Daten und deren Nutzung durch schulische Akteur*innen für die Gestaltung und Entwicklung von Lehr-Lern-Prozessen sowie weiterer schulinterner Prozesse ihre Wirkung als Instrument der Schul- und Unterrichtsentwicklung entfalten. In der Forschungsliteratur lassen sich hierzu verschiedene Folgeaktivitäten im Anschluss an die Durchführung der Vergleichsarbeiten identifizieren. Dabei wird deutlich, dass zum einen zwischen verschiedenen schulinternen Ebenen der Nutzung unterschieden werden kann. Zu differenzieren sind hierbei die individuelle Ebene der einzelnen Lehrkraft und die kollektive Ebene des (Fach-)Kollegiums bzw. der Schule als Ganzes. Zum anderen lassen sich verschiedene Nutzungsaspekte bzw. -schritte identifizieren.

Gemäß der Logik des Zyklusmodells von Helmke und Hosenfeld (2005) bspw. wird zwischen einer Auseinandersetzung mit Ergebnissen durch rezeptive und reflexive Aktivitäten und darauf aufbauenden Folgehandlungen unterschieden. Entsprechend differenzieren die meisten Arbeiten, die das Modell als theoretische Grundlage heranziehen, zwischen diesen Facetten der Ergebnisnutzung. Empirische Untersuchungen verdeutlichen jedoch, dass eine reine Auseinandersetzung mit den rückgemeldeten Daten nicht zwangsläufig zu Anschlussaktivitäten und Unterrichtsentwicklung führt: Koch (2011) ermittelt bspw. nahezu keinen Einfluss der Intensität der Auseinandersetzung mit Ergebnisrückmeldungen als Aspekt von Reflexion auf diverse betrachtete Folgeaktivitäten, konkret Weiterbildungsaktivitäten, Unterrichtsentwicklung, Kooperation und Schulentwicklung. Auch in der Analyse von Wagner et al. (2019) erweist sich die Intensität der Auseinandersetzung mit den Ergebnisrückmeldungen nicht als signifikanter

Prädiktor des Anschlusshandelns, gemessen anhand von Unterrichtsveränderung in Folge der Vergleichsarbeiten.

Auch in Arbeiten, die sich nicht explizit auf einen Zusammenhang zwischen Rezeption bzw. Reflexion und Anschlusshandeln fokussieren, wird deutlich, dass die Ergebnisrückmeldungen zwar häufig zur Kenntnis genommen werden, weitere Aktivitäten aber meist ausbleiben (Ramsteck & Maier, 2015). Dies bestätigt sich bspw. durch die Aussagen vieler Lehrkräfte, die in guten Ergebnissen der Vergleichsarbeiten ihren Unterricht bestätigt sehen und daher weitere Maßnahmen oder Unterrichtsveränderungen als überflüssig erachten (Diemer, 2013; Maier, 2007, 2009a). Bei einem schlechten Abschneiden der eigenen Klasse hingegen besteht u. U. die Gefahr einer Fremdattribution, da schlechte Ergebnisse eher anderen Personen wie Kolleg*innen, Eltern oder den Schüler*innen selbst zugeschrieben werden bzw. auf eine generell eingeschränkte Möglichkeit zur Einflussnahme verwiesen wird, wodurch weitere Maßnahmen wiederum nicht notwendig erscheinen (Schneewind & Kuper, 2009). Obwohl Vergleichsarbeiten somit zwar durchaus für einen Teil der Lehrkräfte diagnostisch relevantes Wissen erzeugen, bleibt die weitere diagnostische Nutzung allerdings hinter den Erwartungen zurück (Maier, 2009b). Viele Lehrkräfte sehen entsprechend aufgrund der Vergleichsarbeiten keinen Anlass für Veränderungen ihrer Unterrichtspraxis (Jäger, 2011).

Hinsichtlich der Auseinandersetzung mit den Ergebnissen finden sich jedoch viele Anhaltspunkte für eine in erster Instanz stattfindende Ergebnisrezeption. In einer Befragung knapp 700 hessischer Lehrkräfte im Nachgang der VERA8-Durchführung im Jahr 2014 geben 90 % der Befragten an, sowohl die Sofortrückmeldungen als auch den Ergebnisbericht zumindest teilweise oder punktuell durchzusehen und zur Kenntnis zu nehmen (Skejic et al., 2015). Einen vergleichbar hohen Grad der individuellen Ergebnisrezeption, von 66 bis nahezu 100 % je nach Untersuchung, beschreibt auch Dedering (2011) in einem Literaturreview und ermittelt zusätzlich eine Abhängigkeit von Schulform und Unterrichtsfach. Hinweise für eine auf der ersten Rezeption der Ergebnisse aufbauende formative Unterrichtsevaluation einschließlich daraus resultierender Schul- und Unterrichtsentwicklungsprozesse gibt es dagegen nur in einem deutlich geringeren Ausmaß (Skejic et al., 2015). Auch in anderen Survey-Studien verneinen bis zu 50 % der Lehrkräfte die Frage, ob VERA zu einer kritischen Reflexion des eigenen Unterrichts und zur Ableitung von Weiterentwicklungsmaßnahmen beiträgt (Maier, 2009d).

Das Interesse der Lehrkräfte bei der Rezeption der Ergebnisrückmeldungen einhergehend mit der größten Bereitschaft einer weiteren Auseinandersetzung richtet sich primär auf die

individuelle Ebene der einzelnen Schüler*innen und der eigenen Schulklasse (Kuper & Diemer, 2012; Skejic et al., 2015). Der Fokus der Rezeption liegt häufig auf dem Gesamtergebnis inklusive der Stärken und Schwächen der eigenen Schüler*innen bzw. Klasse (Dedering, 2011; Maier, 2009b). Insbesondere das Abschneiden der Schüler*innen bei einzelnen Aufgaben bzw. der Vergleich bestimmter Aufgabenbereiche, Vergleiche mit Parallelklassen und Landesdurchschnittswerten und auch das Gesamtergebnis der Schule werden hierbei von Lehrkräften als relevante Informationen aufgeführt (Diemer, 2013). Bis zu 70 % der Lehrkräfte geben in diesem Zusammenhang in einer Befragung an, sich für das Abschneiden der eigenen Klasse im Vergleich zu anderen Referenzklassen zu interessieren (Skejic et al., 2015). Ein an Kompetenzstufen orientierter Umgang mit den rückgemeldeten Daten ist hingegen kaum erkennbar. Kompetenzstufenbeschreibungen und didaktische Handreichungen werden nach Aussagen von Lehrkräften bei der Interpretation der Leistungsdaten eher nicht berücksichtigt (Kuper, Maier, Graf, Muslic & Ramsteck, 2016).

Der grundlegenden Intention der Vergleichsarbeiten nach sollte die Auseinandersetzung mit den Ergebnissen auf verschiedenen Ebenen des schulischen Systems stattfinden. In der Praxis kann jedoch insgesamt von einer flächendeckenden Nutzung keine Rede sein. Wenn die Ergebnisse aus Vergleichsarbeiten genutzt werden, dann meist mit Fokus auf einzelne Maßnahmen (Wurster et al., 2017). Nur selten werden aus externen Evaluationsverfahren wie den Vergleichsarbeiten in größerem Ausmaß klar definierte unmittelbare Konsequenzen abgeleitet, und der Einfluss auf das Handeln schulischer Akteur*innen erweist sich als eher gering (Dedering, 2011). Untersuchungen zeigen, dass bei den Maßnahmen, die aus der Rezeption der Ergebnisrückmeldungen abgeleitet werden, Aktivitäten auf Unterrichtsebene in der Praxis klar im Vordergrund stehen, während die VERA-Ergebnisse auf Schulebene eher selten weitere Verwendung finden (z.B. Bach et al., 2014; Koch, Groß Ophoff, Hosenfeld & Helmke, 2006; Wurster, Bach et al., 2016).

Hinsichtlich der individuellen Nutzung lassen sich verschiedene Nutzungsaspekte bzw. Folgeaktivitäten identifizieren. Die intensivste weitere Nutzung der Rückmeldungen lässt sich im Bereich der Unterrichtsinhalte ausmachen (Diemer, 2013; Groß Ophoff, Koch, Helmke & Hosenfeld, 2006). Zu den in verschiedensten Beiträgen meistgenannten Maßnahmen zählt die Ausrichtung des weiteren Unterrichts auf Unterrichtsinhalte, die sich im Test als problematisch herausgestellt haben (Diemer, 2013), einschließlich einer verstärkten inhaltlichen Schwerpunktsetzung auf Themenbereiche, die in den Vergleichsarbeiten abgeprüft werden, was zu einer Verengung bzw. Verlagerung von Unterrichtsinhalten führen kann (Diemer, 2013; Jäger,

2011). Der Fokus liegt hierbei häufig auf Wiederholung, Üben und Vertiefung einzelner Themengebiete, in denen durch den Test Schwächen der Schüler*innen offengelegt wurden (Dedering, 2011; Koch et al., 2006; Maier, 2007). Auch werden mitunter bestimmte als innovativ empfundene Aufgabenformate, die im eigenen Unterricht verstärkt geübt werden sollen, von den Lehrkräften übernommen (Ramsteck & Maier, 2015; Wurster, Bach et al., 2016) sowie an die Testaufgaben angelehnte Unterrichtsmaterialien entwickelt (Koch et al., 2006). In diesem Kontext erfolgt auch vereinzelt das Einüben des Lesens und Verstehens von Arbeitsanweisungen mit den Schüler*innen (Kuper et al., 2016) und die Vermittlung bestimmter Arbeitstechniken (Maier, 2009a). Vereinzelt konstatieren Lehrkräfte auch die Absicht, Methoden zu überdenken bzw. Lehr-Lern-Methoden einzuführen (Jäger, 2011; Koch et al., 2006).

Weitere individuelle Nutzungsvarianten beziehen sich bspw. auf Maßnahmen der Einzelförderung (Wurster et al., 2017). Basierend auf einer Bestimmung und Analyse von Stärken und Schwächen einzelner Schüler*innen (Wurster, Bach et al., 2016), gibt es durchaus Lehrkräfte, die die VERA-Ergebnisse als Anlass für eine stärkere Differenzierung des Unterrichts sehen und die Rückmeldungen zu einer individuellen Förderung nutzen (Diemer, 2013; Ramsteck & Maier, 2015). Vereinzelt werden auch längerfristig vorbereitende Maßnahmen, wie Probetests und spezieller Förderunterricht, als Konsequenz aus den VERA-Rückmeldungen angeführt (Diemer, 2013).

Darüber hinaus erweist sich die Kommunikation mit anderen Interessensgruppen im Hinblick auf die Ergebnisse der Vergleichsarbeiten als zu berücksichtigende Maßnahme infolge der Testungen. Hierbei stellt sich die Kommunikation mit den Eltern über das Abschneiden der Schüler*innen als verhältnismäßig häufige Aktivität im Anschluss an die Vergleichsarbeiten heraus (Skejic et al., 2015). Als Maßnahme der Elternberatung erfolgt u. U. auch eine Information der Eltern über Ergebnisse und Hintergründe von VERA, meist handelt es sich jedoch um eine reine Inkenntnissetzung bzgl. der Teilnahme und des Abschneidens der Schüler*innen (Ramsteck & Maier, 2015). Vereinzelt werden die Ergebnisse auch in Elterngesprächen als eine Art objektives Kriterium zur Rechtfertigung von Entscheidungen in der Notengebung oder für Schulempfehlungen herangezogen (Maier, 2009b). Gelegentlich finden auch informelle, teils jedoch systematische Gespräche mit einzelnen Schüler*innen zu den Ergebnissen der Vergleichsarbeiten statt (Jäger, 2011).

Ein weiterer Aspekt der Kommunikation betrifft den schulinternen Austausch im Kollegium. Diesbezüglich lässt sich festhalten, dass, wenn überhaupt, informelle Gespräche über

Ergebnisse oder technische Aspekte der Durchführung mit ausgewählten Kolleg*innen stattfinden (Jäger, 2011; Maier, 2009a; Skejic et al., 2015). Eine lerngruppenübergreifende Besprechung der Ergebnisse findet maximal in den Fachkonferenzen statt, in Gesamtkonferenzen werden die Vergleichsarbeiten i. d. R. kaum thematisiert (Dedering, 2011; Maier et al., 2011).

In der Konsequenz eines Mangels an systematischer schulinterner Kommunikation über VERA werden tendenziell auch nur selten Maßnahmen auf Schulebene abgeleitet (Maier, 2009a). Systematische Kooperationen oder aus den Vergleichsarbeiten resultierende Maßnahmen der Personal- und Schulentwicklung sind somit eher selten auszumachen (Koch et al., 2006). Zu diesen zumindest vereinzelt umgesetzten und im Sinne einer durch die Vergleichsarbeiten angestoßenen Schulentwicklung wünschenswerten Maßnahmen zählen u. a. eine bereits angesprochene systematische Kommunikation im Kollegium einschließlich einer wechselseitigen Unterstützung. Diese betrifft insbesondere Fragen der Durchführung sowie eine gemeinsame Korrektur, welche zumindest gelegentlich stattfindet. Unterstützungsmaßnahmen durch die Schulleitungen finden sich dagegen eher selten, maximal in vereinzelten schulischen Fortbildungen oder Hospitationen bei Lehrkräften besonders schlecht abschneidender Klassen (Ramsteck & Maier, 2015). Sporadisch zeigen sich Maßnahmen der Personalentwicklung wie die Schaffung von Stellen zur Koordination von Vergleichsarbeiten oder der Entwicklung von Förderkonzepten (Bach et al., 2014; Ramsteck & Maier, 2015). Weitere denkbare und zuweilen auftretende Maßnahmen sind die Veränderung von Lehrplänen oder die Einführung einheitlicher Bewertungsmaßstäbe in Folge der Vergleichsarbeiten (Diemer, 2013). Eine Kontrolle der Umsetzung von Maßnahmen zur Schul- und Unterrichtsentwicklung in Folge von Vergleichsarbeiten findet auf Schulebene ebenfalls kaum statt. Die Verantwortung hierfür wird durch die Schulaufsicht an die Schulen delegiert, wo jedoch höchstens vereinzelt Gespräche zwischen Schulleitung und einzelnen Lehrkräften, deren Klassen besonders schlecht abschneiden, stattfinden (Ramsteck & Maier, 2015).

Insgesamt belegen Untersuchungen zur Weiterarbeit mit Daten aus Vergleichsarbeiten, dass diese eher selten genutzt werden und die Ergebnisnutzung wenig zufriedenstellend erscheint (Ercan, Hartmann, Richter, Kuschel & Gräsel, 2021; Kühle & Peek, 2007). In Befragungen liegen die meisten erfragten Maßnahmen und Folgeaktivitäten häufig unter dem theoretischen Mittelwert (z.B. Ercan et al., 2021; Kühle & Peek, 2007; Richter et al., 2014). Generell wird deutlich, dass die Nutzung der VERA-Daten eher auf unterrichtsnahe Maßnahmen, wie eine gezielte Wiederholung von Unterrichtsstoff, Aufgabenentwicklung oder individuelle Einzelförderung, abzielt (Muslic, 2017; Wurster, Bach et al., 2016). Selten erfolgt eine Nutzung von

Evaluationsdaten für strategische Entscheidungen wie Personalentwicklung und Planung von Schulprogrammen (Bach et al., 2014; Wurster, Bach et al., 2016). Rückmeldedaten aus Vergleichsarbeiten werden somit, wenn überhaupt, in stärkerem Ausmaß zur Unterrichtsentwicklung als zur Schulentwicklung genutzt (Bach et al., 2014).

Für die unzureichende Nutzung der Daten aus Vergleichsarbeiten gibt es in der Literatur verschiedene Erklärungen, die an dieser Stelle nur knapp angerissen werden, jedoch teils in den anderen Abschnitten dieses Kapitels 2.2.2 als Einflussfaktoren bzw. Wahrnehmungsfaktoren ausführlicher behandelt werden. Es findet sich immer wieder der Hinweis, dass die Nutzung bzw. Nutzungsintention sowohl auf Ebene der Lehrkräfte als auch auf Schulebene von der Einschätzung der Nützlichkeit der Vergleichsarbeiten abhängt (z.B. Dedering, 2011; Vogel, 2020; Wurster, Bach et al., 2016). Als weitere die Ergebnisnutzung begünstigende Einflüsse erweisen sich eine positive Einstellung und eine ausgeprägte Akzeptanz der schulischen Akteur*innen (Muslic, 2017). Wichtiger Bedingungsfaktor für die Ergebnisnutzung ist zudem die Verständlichkeit zunächst des gesamten Verfahrens und insbesondere der Ergebnisrückmeldungen. Das Vorhandensein unterstützender Strukturen wirkt zusätzlich förderlich auf die Ergebnisnutzung (Jäger, 2011; Schneewind & Kuper, 2009), ebenso wie ein positives Schulklima (Dedering, 2011). Jedoch hängt die Weiterarbeit mit den Ergebnissen auch von der individuellen Bereitschaft und den vorhandenen Ressourcen der Lehrkräfte ab (Diemer, 2013; Schneewind & Kuper, 2009). In diesem Zusammenhang konstatieren viele Lehrkräfte auch ein grundlegendes Desinteresse, bspw. aufgrund von fehlender Zeit oder mangelnder wahrgenommener Relevanz von Vergleichsarbeiten, da wichtigere Probleme in ihren Klassen bestehen (Jäger, 2011). Vergleichsarbeiten werden demnach häufig von Lehrkräften hauptsächlich als zeitintensive Mehrbelastung gesehen (Ramsteck & Maier, 2015). Als zusätzliches Hemmnis der Nutzung stellt sich die Wahrnehmung von VERA als wenig valides Testinstrument heraus, in welchem primär Randthemen und keine zentralen Unterrichtsinhalte abgeprüft werden (Maier, 2009b; Ramsteck & Maier, 2015). Weiterhin werden Zweifel an der Aussagekraft der Testergebnisse dadurch begründet, dass Schüler*innen Aufgaben nicht sorgfältig bearbeiten, was zu einer Verfälschung der Ergebnisse führt, ebenso wie die teils als ungerecht erachtete dichotome Punktevergabe ohne Berücksichtigung von Teilpunkten (Maier, 2009b).

Zusammenfassend lässt sich hinsichtlich der Nutzung der VERA-Ergebnisse resümieren, dass diese zwar häufig zumindest von den Lehrkräften zur Kenntnis genommen werden, eine vertiefte Auseinandersetzung und weitere Verwendung jedoch deutlich seltener stattzufinden scheint. Wenn eine Weiterarbeit erfolgt, dann i. d. R. eher unterrichtsnah und auf einzelne

Maßnahmen, wie die Übernahme von Aufgaben oder Einzelförderung von Schüler*innen fokussiert. Maßnahmen und Aktivitäten auf der kollektiven Schulebene zeigen sich dagegen eher in Einzelfällen. Die Aufarbeitung der entsprechenden Literatur legt somit die Vermutung nahe, dass Vergleichsarbeiten in der Praxis – falls eine weitere Verwendung stattfindet – eher als Instrument der Unterrichtsentwicklung, denn als Schulentwicklungsinstrument genutzt werden.

2.2.2.5. Zeitliche Belastung – Aufwand-Nutzen

Ein prägender Faktor der Wahrnehmung von Lehrkräften hinsichtlich VERA ist der mit der Durchführung, Auswertung und Weiterarbeit verbundene Zeitaufwand; eine Problematik, die vor allem in Interviewstudien mehrfach thematisiert wird. Vergleichsarbeiten werden demnach häufig als zeitintensive Zusatzbelastung angesehen, was wiederum einer weiteren Nutzung der Rückmeldungen abträglich ist (Ramsteck & Maier, 2015). Lehrkräfte sowie Schulleitungen bemängeln dabei besonders den zeitlichen Aufwand durch die Dateneingabe (Kuper et al., 2016), der sich zusammen mit dem Aufwand für Durchführung und Auswertung für die Ablehnung des Instruments Vergleichsarbeiten mitverantwortlich erweist (Skejic et al., 2015). Der mit der Durchführung der Vergleichsarbeiten verbundene Aufwand variiert jedoch zwischen den eingesetzten Testheften. Eine Analyse von Selbsteinschätzungen der Auswertungszeit für Korrektur und Dateneingabe ergibt, verglichen mit den sprachlichen Fächern Deutsch und Englisch, in Mathematik eine signifikant geringere Auswertungszeit. Auch die Bewertung der Angemessenheit der Auswertungszeit fällt in den Fächern Deutsch und Englisch niedriger aus als in Mathematik (Vogel, 2020).

Studien belegen den Zusammenhang von wahrgenommener zeitlicher Belastung und anderen Wahrnehmungsfaktoren sowie der Weiterarbeit mit den VERA-Rückmeldungen. Einer ausführlichen und systematischen Auseinandersetzung mit den Rückmeldungen steht häufig im Wege, dass zeitliche und personelle Ressourcen schon zuvor durch die zusätzliche große Arbeitsbelastung durch Korrektur der Lernstandserhebungen und anschließende Datenübermittlung gebunden werden (Diemer, 2013). Zusätzlich finden sich Hinweise auf einen Zusammenhang der wahrgenommenen Belastung durch Vergleichsarbeiten und deren Akzeptanz (Vogel et al., 2016). Interviews mit Lehrkräften belegen, dass ein hoher Organisationsaufwand, der als belastend empfunden wird, negativ auf die Akzeptanz wirkt und somit indirekt eine ausbleibende Datenrezeption bedingt (Muslic, 2017).

Des Weiteren finden sich empirische Belege für die Vorhersagekraft der Auswertungszeit auf die Nützlichkeitswahrnehmung von Lehrkräften (Vogel, 2020) und mittlere bis hohe negative Korrelationen zwischen Aufwand und Nützlichkeit (Wurster & Richter, 2016). Dabei erweist sich die Balance zwischen Aufwand und Ertrag als wichtige Voraussetzung für die Nutzung von Lernstandserhebungen, im Sinne einer Kosten-Nutzen-Abwägung. In dieser Bilanzierung schlägt der zeitliche Aufwand durch Korrektur und Ergebniseingabe auf der Kostenseite zubuche, während der wahrgenommene Nutzen der Ertragsseite zuzurechnen ist (Vogel, 2020).

Auch der folgende kurze Literaturüberblick spiegelt die Bedeutung eines ausgeglichenen Kosten-Nutzen-Verhältnisses wider: In einer bereits 2005 durchgeführten Onlinebefragung identifizieren Bonsen et al. (2006) die Balance von Aufwand und Ertrag als eine Gelingensbedingung von Lernstandserhebungen. Auch Vogel et al. (2016) arbeiten die Balance zwischen Aufwand und Ertrag, bezeichnet als Praktikabilität, als bedeutendes Kriterium eines erfolgreichen Einsatzes von Lernstandserhebungen heraus. Dabei wird die wahrgenommene Belastung durch die Lernstandserhebungen als ein entscheidender Prädiktor dieser Balance identifiziert.

Wie verschiedene Untersuchungen zeigen, kommt es in der Praxis häufig zu einem Missverhältnis zwischen empfundenem Aufwand und Ertrag der Vergleichsarbeiten, sodass die Balance in der Wahrnehmung der Lehrkräfte gestört ist. Schlägt die Abwägung von Aufwand und Nutzen zuungunsten des Nutzens aus, kann dies als eine der relevantesten Ursachen für eine fehlende Weiterarbeit mit den Ergebnissen gesehen werden. Alle befragten schulischen Akteur*innen – Schulleitungen, Fachbereichsleitungen und Lehrkräfte - weisen auf eine Diskrepanz zwischen Aufwand und Nutzen der Vergleichsarbeiten für die Unterrichts- und Schulentwicklung hin (Muslic, 2017). In den Interviewstudien von Demski (2019a, 2019b) betonen Lehrkräfte den durch die Durchführung und insbesondere die Korrektur entstehenden Mehraufwand der Vergleichsarbeiten, der keinen Ausgleich finden würde, sowie das generell ungünstige Verhältnis von Aufwand und Ertrag bzw. Nutzen. Kuhn (2014) konstatiert generell ein in der Schulpraxis wahrgenommenes ungünstiges Verhältnis von Aufwand und Ertrag der Vergleichsarbeiten, während Kuper et al. (2016) anmerken, dass gerade Fachbereichsleitungen den Aufwand der Vergleichsarbeiten hinsichtlich der benötigten Zeit vor dem Hintergrund des daraus resultierenden Nutzens im Sinne eines geringen Erkenntnisgewinns kritisieren. Wurster und Bach et al. (2016) finden bei einer Befragung verschiedener schulischer Akteur*innen heraus, dass Lehrkräfte Vergleichsarbeiten als eher weniger nützlich und eher aufwendig einschätzen, eine Beurteilung, die, ebenso wie die allgemeine Einstellung, bei Fachkonferenz- bzw. Schulleitungen etwas positiver ausfällt. Generell fällt die Aufwand-Nutzen-Bewertung der

Vergleichsarbeiten umso ungünstiger aus, je direkter die Befragten an der Durchführung beteiligt sind.

2.2.2.6. Weitere Wahrnehmungsaspekte und Gründe für (Nicht-)Nutzung

Neben den in den vorangegangenen Abschnitten bereits ausführlich beschriebenen Wahrnehmungs- und Bewertungsaspekten der Vergleichsarbeiten finden sich in der Literatur weitere Faktoren, welche die Wahrnehmung und Nutzung der Vergleichsarbeiten beeinflussen. Diese werden teils ausführlicher, teils nur wenig detailliert untersucht und sind für diese Arbeit zwar nur zweitrangig, werden jedoch aus Gründen der Vollständigkeit in diesem Kapitel zumindest knapp dargestellt.

Wahrgenommene Funktion der Vergleichsarbeiten, Kontrollempfinden und Misstrauen

Zunächst spielt die Funktionswahrnehmung der Vergleichsarbeiten eine Rolle bei deren Beurteilung durch Lehrkräfte. In deren Wahrnehmung lassen sich demnach verschiedene intendierte und teils auch nicht intendierte Funktionen erkennen. In der Literatur wird dabei hauptsächlich der mit den Vergleichsarbeiten auch intendierten Entwicklungsfunktion eine eher unerwünschte Kontrollfunktion gegenübergestellt, die aus Sicht einiger schulischer Akteur*innen nicht unerheblich scheint (z.B. Jäger, 2011; Maier, 2009d; Richter & Böhme, 2014). Diese wahrgenommene Kontrollfunktion stützt sich auf die Befürchtung, Vergleichsarbeiten seien insbesondere als Instrument zur Überprüfung von Lehrkräften bzw. deren Unterrichtsqualität implementiert worden. Insbesondere mit bzw. in den ersten Jahren nach Einführung der Vergleichsarbeiten kristallisierte sich bei vielen Lehrkräften die Befürchtung zunehmender Kontrolle durch die Durchführung derartiger Leistungstests heraus. In einer Befragung in Nordrhein-Westfalen im Schuljahr 2004/2005 äußern zwischen 60 % und 70 % der befragten Lehrkräfte die Vermutung, durch die landesweiten Lernstandserhebungen stärkerer Kontrolle ausgesetzt zu sein. Gleichzeitig zeigen sich jedoch auch positive Einstellungen und viele Lehrkräfte sehen in den Vergleichsarbeiten Entwicklungspotenziale für Schule und Unterricht (van Ackeren & Bellenberg, 2004). Auch in späteren quantitativen Befragungen zeigen sich sowohl zu verschiedenen Aspekten einer Schul- und Unterrichtsentwicklungsfunktion von Vergleichsarbeiten als auch zu diversen Kontrollfunktionsstatements Zustimmungstendenzen. Ein eindeutiger Trend ist dabei schwer auszumachen, insgesamt scheint jedoch im Vergleich die Wahrnehmung von Vergleichsarbeiten als ein Instrument mit Entwicklungsfunktionen zu überwiegen (Richter et al., 2014; Richter & Böhme, 2014).

Wirklich bedeutsam wird die Funktionswahrnehmung der Vergleichsarbeiten erst durch die daraus resultierenden Konsequenzen für die Weiterarbeit mit den Ergebnisrückmeldungen. Einige Untersuchungen kommen in diesem Zusammenhang zu dem Schluss, dass die Wahrnehmung der Funktion den Umgang mit den Ergebnissen beeinflusst (Kühle & Peek, 2007; Richter et al., 2014). Eine Untersuchung von Richter et al. (2014) ermittelt bspw. einen positiven Zusammenhang zwischen wahrgenommener Unterrichtsentwicklungsfunktion und dem Selbstbericht einer Kompetenzorientierung im Unterricht sowie einer zunehmenden Differenzierung des Unterrichts. Ein vergleichbarer Zusammenhang zeigt sich nicht bei einer Wahrnehmung als Kontrollinstrument. Hinsichtlich einer möglichen Verengung des Lehrplans in Folge der Vergleichsarbeiten, als tendenziell negativ konnotierte Konsequenz, erweisen sich sowohl eine Wahrnehmung als Kontrollinstrument als auch die einer Entwicklungsfunktion als prädiktiv. Dieser Zusammenhang zwischen wahrgenommener Entwicklungsfunktion und Verengung des Lehrplans erscheint auf den ersten Blick wenig plausibel und bedarf auch aus Sicht der Autor*innen weiterer Untersuchungen. Zusätzlich liefern die Ergebnisse der Arbeit Hinweise darauf, dass Lehrkräfte eher ihren Unterricht verändern, wenn sie VERA als förderlich für die Unterrichtsentwicklung wahrnehmen und den Vergleichsarbeiten diagnostische Informationen entnehmen können.

Ein durch VERA hervorgerufenes Gefühl der Kontrolle bzw. des Kontrolliertwerdens steht im Verdacht, mögliche weitere negative oder unerwünschte Verhaltensweisen hervorzurufen, die sich auf die Reputation der Vergleichsarbeiten und deren Wahrnehmung in der Öffentlichkeit bzw. durch andere Lehrkräfte auswirken können. Neben einer befürchteten Verengung des Lehrplans wären bspw. die Vermutung eines zunehmenden Teaching to the Test zu nennen, also eine gezielte Vorbereitung auf die Vergleichsarbeiten, um ein gutes Abschneiden der eigenen Schüler*innen zu fördern. Diese Befürchtung wird auch durch empirische Befunde untermauert: Rund 60 % bis 80 % der befragten Lehrkräfte verspüren einen gewissen Druck, gute Testergebnisse erzielen zu müssen (Bellmann et al., 2016). Demzufolge erscheint es durchaus plausibel, dass in einer anderen Lehrkräftebefragung bundeslandabhängig zwischen 39 % und 49 % der Befragten angeben, nach Erhalt der Testhefte vor dem Testtag bereits Aufgaben mit ihren Schüler*innen zu üben. Aus Sicht der Autor*innen geschieht dies, um schlechte Testergebnisse und die damit einhergehende Verantwortlichkeit zu vermeiden (C. Thiel et al., 2017). Diese Vermutung eines teils unfairen Vorbereitungsverhaltens von Kolleg*innen äußert sich auch in der Interviewstudie mit Lehrkräften von Jäger (2011). Insgesamt scheint sich die Befürchtung einer übermäßigen und auch unfairen Testvorbereitung jedoch eher weniger bzw.

eher in Einzelfällen zu bestätigen (Richter et al., 2014; Skejic et al., 2015; Wacker & Kramer, 2012), dennoch finden sich in Forschungsarbeiten durchaus Hinweise auf ein verstärktes Konkurrenzdenken zwischen Lehrkräften bezüglich des Abschneidens der eigenen Schüler*innen (Demski, 2019b).

Befeuert wird die Thematik vor allem durch Erkenntnisse angloamerikanischer Studien im Kontext von high-stakes Testing. Kuper et al. (2016) verweisen bspw. auf Befunde in amerikanischen Forschungsarbeiten zu eindeutig belegtem unseriösen Verhalten von Lehrkräften zur Erhöhung von Testwerten. Die Autor*innen merken an, dass ein solches Vorgehen zwar in Deutschland nicht belegt ist, es jedoch durch milde Bewertungen von Lehrkräften oder gezieltes Üben von Testaufgaben durchaus Hinweise in diese Richtung gibt. Gerade wenn Lehrkräfte Vergleichsarbeiten als eine Art high-stakes Instrument der Fremdevaluation und -kontrolle wahrnehmen, könnten sie dazu verleitet werden, die Ergebnisse zu „schönen“ und die Testvalidität dadurch zu gefährden. Jedoch gelangen die Autor*innen zu dem Schluss, dass es sich in der Praxis eher um Einzelfälle handelt, die nicht im großen Stil auftreten sollten (Leutner et al., 2008).

Dennoch stehen, gerade vor dem Hintergrund der Durchführungsmodalitäten der Vergleichsarbeiten, Durchführungs- und Auswertungsobjektivität grundsätzlich in der Kritik (z.B. Bonsen et al., 2006; Jäger, 2011; C. Thiel et al., 2017). Da die Testleitung der Vergleichsarbeiten, ebenso wie die spätere Auswertung in der Verantwortung der Fachlehrkräfte liegt, besteht zumindest die Möglichkeit, dass Unregelmäßigkeiten in der Testdurchführung, wie bspw. Hilfestellungen, Tipps oder das Tolerieren von Abschreiben, auftreten (Bonsen et al., 2006). Die Lehrkräftebefragung von Skejic et al. (2015) zeigt, dass diese Vermutung nicht ganz unbegründet scheint, da immerhin 22 % der Befragten angeben, ihre Schüler*innen während der Testdurchführung zumindest teilweise unterstützt zu haben, jedoch mit der Absicht, die Schüler*innen während des Tests nicht zu demotivieren. Die Untersuchung von Bellmann et al. (2016) ermittelt für alle Bundesländer ähnliche Werte bei der Frage nach der Unterstützung der Schüler*innen während des Tests.

Im Hinblick auf die Auswertung der Vergleichsarbeiten werden mehrere Aspekte von Lehrkräften kritisch gesehen. Zum einen wird die Punktevergabe bemängelt, da hierbei häufig keine Teilleistungen berücksichtigt werden, was in der Schule sonst gängige Praxis darstellt (Jäger, 2011; Spoden et al., 2014). Darüber hinaus stellt sich in Interviews heraus, dass sich nicht alle Lehrkräfte an die Bewertungsvorgaben halten und auch zugunsten der Schüler*innen bspw.

Teilpunkte vergeben (Jäger, 2011). Weitere Evidenz für eine mangelnde Auswertungsobjektivität und ein Abweichen von Lehrkräften von den normierten Auswertungsvorlagen finden Spoden et al. (2014) und unterscheiden dabei zwischen unsystematischen kognitiven und systematischen motivationalen Beurteilungsfehlern. Den Autor*innen zufolge sollten systematische Bewertungsfehler durch die eindeutigen Auswertungsanleitungen weitestgehend ausgeschlossen sein, motivational bedingte Abweichungen der Bewertung scheinen andererseits durch die Wahrnehmung der Vergleichsarbeiten als Kontrollinstrument befördert.

Auch Koch und Hosenfeld (2013) untersuchen anhand unabhängiger Zweit-Beurteilungen von VERA-Testheften die Auswertungsobjektivität bei Vergleichsarbeiten. Die Analyse basiert auf den VERA3-Tests zum Leseverstehen der Jahre 2009 bis 2012. Datengrundlage bilden dabei Schüler*innentesthefte, die mit Hilfe einer geschichteten Zufallsstichprobe ausgewählt und einer Zweitkorrektur durch geschulte Kodierer*innen unterzogen wurden. Die Auswertung der Zweitkorrektur im Vergleich zur Korrektur der Lehrkräfte zeigt, dass es bei 3.4 % der Antworten zu Abweichungen kommt, wobei insgesamt eher eine Tendenz zur Überschätzung der Schüler*innenleistungen zutage tritt. Bestimmte Aufgabeneigenschaften beeinflussen wiederum die Ausprägung der Abweichungen, bspw. zeigen sich bei Multiple-Choice-Aufgaben deutlich seltener Abweichungen als bei alternativen Antwortformaten wie offenen Aufgaben. Die Autor*innen finden somit durchaus auch empirische Belege für die Vermutung einer nicht vollständig vorgabenkonformen Korrektur durch Lehrkräfte. Die gefundenen Abweichungen werden dabei zwar als gering eingestuft, dennoch resümieren die Autor*innen, dass je nach Charakter der Fehlurteile durchaus Reliabilität und Validität des Testverfahrens beeinflusst werden könnten.

Somit finden sich zwar einzelne Belege für ein unfaires bzw. einem nicht dem Standard entsprechenden Verhalten von Lehrkräften, jedoch scheinen diese in der Praxis, zumindest gerade basierend auf Selbstauskünften über das eigene Verhalten, eher Ausnahmen als die Regel darzustellen. Vogel (2020) identifiziert in diesem Zusammenhang das Vorhandensein von Akzeptanz bei Lehrkräften als wichtige Voraussetzung einer verzerrungsfreien Durchführung und Nutzung des Instruments. Auch wenn deviantes Verhalten von Lehrkräften eher selten zu belegen scheint, bedeutet dies nicht, dass nicht dennoch die Wahrnehmung von Lehrkräften im Hinblick auf das Verhalten von Kolleg*innen und dadurch auch auf das Instrument VERA durchaus verzerrt sein kann und ein gewisses Misstrauen herrscht (z.B. Zuber, 2019). Wie bspw. Demski (2019b) zeigt, werden durchaus von Lehrkräften Zweifel an Validität und Reliabilität des Instruments Vergleichsarbeiten und dessen Ergebnissen geäußert. Die

Konsequenzen von Misstrauen ggü. der Güte des Testinstruments, der Testdurchführung etc. für die Praxis der Vergleichsarbeiten und konkret die Nutzung der Ergebnisrückmeldungen, bleiben jedoch unklar.

Wissen, Verständnis und Verhaltenskontrolle

Weitere Aspekte, die die Wahrnehmung der Vergleichsarbeiten beeinflussen könnten und zumindest in einigen Arbeiten wenigstens oberflächlich beleuchtet wurden, betreffen die Informiertheit der Lehrkräfte über die Vergleichsarbeiten sowie deren Verständnis im Umgang mit den rückgemeldeten Daten. Hierbei lassen sich Parallelen zu den Einstellungstheorien aus Kapitel 2.1.2 ziehen, konkret zur Theory of Planned Behavior (TPB). In dieser stellt die wahrgenommene Verhaltenskontrolle einen wichtigen Verhaltensprädiktor dar und definiert sich als das Gefühl einer Person, auf alle für eine Handlung notwendigen Ressourcen leicht zugreifen zu können, sowohl die eigene Kompetenz als auch Umweltressourcen betreffend. Auf die Vergleichsarbeiten übertragen stellt sich in der Arbeit von Zuber (2019) die Bewertung der Qualität und Verständlichkeit der zur Verfügung gestellten Informationsmaterialien als ein Aspekt wahrgenommener Verhaltenskontrolle heraus. Weitere Faktoren können anhand der Aussagen der Lehrkräfte nicht als Elemente von Verhaltenskontrolle identifiziert werden. Wie in der TPB postuliert, rekonstruiert Zuber (2019) basierend auf der qualitativen Auswertung von Interviews mit Lehrkräften auch einen positiven Einfluss der wahrgenommenen Verhaltenskontrolle auf das Umsetzungshandeln der Lehrkräfte.

Insgesamt scheint zunächst die Informiertheit der Lehrkräfte durchwachsen. Generell erweisen sich Lehrkräfte im Gesamten als gut über VERA informiert, besonders im Hinblick auf Durchführung und Nutzungsvorschriften. Bei der Informiertheit über die Ziele der Vergleichsarbeiten gibt es hingegen noch Nachbesserungsbedarf (Wurster, Bach et al., 2016). In der Arbeit von Skejic et al. (2015) fühlen sich sogar nur rund 50 % der Befragten gut auf VERA vorbereitet, wohingegen Jäger (2011) nahezu keine Probleme bei der Durchführung identifiziert.

Auch die Verständlichkeit der Ergebnisrückmeldungen erweist sich aus Sicht der Lehrkräfte, abhängig von der jeweiligen Untersuchung, als durchmischt. Es zeigen sich dahingehend durchaus häufig positive Einschätzungen von Lehrkräften, wobei in Mathematik deutlich weniger Verständnisschwierigkeiten auftreten als in den sprachlichen Fächern (Dedering, 2011). Derartige Indizien, dass Mathematiklehrkräfte generell besser mit den Rückmeldungen zurechtkommen als Lehrkräfte in sprachlichen Fächern, finden sich auch bei Diemer (2013), der auch

vereinzelte Probleme bei der Interpretation von Grafiken offenlegt. Auch Demski (2019b) merkt an, dass ein Teil der Lehrkräfte vom Informationsgehalt der Rückmeldungen überfordert scheint und Schwierigkeiten im Ableiten konkreter Anschlusshandlungen hat oder nur schwer einen Bezug zum eigenen Unterricht herstellen kann. In diesem Kontext resümiert Jäger (2011), dass Verständlichkeit alleine auch keine hinreichende Voraussetzung für ein Anschlusshandeln darstellt, eine Vermutung, für die sich bei Kühle und Peek (2007) auch empirische Belege finden. Die Autoren erachten zwar zunächst die Verständlichkeit bzw. das Verständnis der Rückmeldungen als essenzielle Voraussetzung einer produktiven Ergebnisnutzung, der Zusammenhang erweist sich jedoch letztendlich in einer empirischen Überprüfung als nicht signifikant. Koch (2011) dagegen findet einen positiven Effekt der Verständlichkeit der Ergebnisrückmeldungen, als Aspekt von Rezeption, auf die Reflexion der Ergebnisse. Je verständlicher Lehrkräfte Rückmeldungen wahrnehmen, desto intensiver setzen sie sich demnach mit diesen auseinander, und desto nützlicher werden sie außerdem wahrgenommen.

Innerschulischer Umgang mit VERA, Rolle der Schulleitung und Subjektive Norm

Zum Abschluss dieses Kapitels werden schulinterne Einflüsse auf die Nutzung der Daten aus Vergleichsarbeiten beleuchtet. Hierzu gibt es u. a. einige Erkenntnisse hinsichtlich der Wirkung kollegialer Strukturen oder der Einflussnahme der Schulleitungen. So kann wiederum die Brücke zur Theory of Planned Behavior (TPB) bzw. zur Theory of Reasoned Action (TRA) geschlagen werden. Hierbei kann der Einfluss von Schulleitung sowie sonstigen innerschulischen Strukturen, Kooperationen etc. als eine Facette von subjektiver Norm interpretiert werden, die die individuellen Einflüsse des sozialen Umfeldes repräsentiert, also im Fall der Vergleichsarbeiten das schulische Umfeld einer Lehrkraft. Zunächst losgelöst vom Kontext der Vergleichsarbeiten erweisen sich bestimmte schulische Strukturen als prädiktiv für die Evidenzorientierung von Lehrkräften. Demnach begünstigen elaborierte Kommunikations- und Informationsverteilungsstrukturen, interne (z. B. kollegialer Austausch, Fachkonferenzen) und externe (z. B. Austausch zwischen Schulen) Kooperationsstrukturen sowie ausgeprägte Partizipationsmöglichkeiten die allgemeine Evidenzorientierung von Lehrkräften (Zlatkin-Troitschanskaia, Förster, Preuße & Mater, 2016). Dabei erweisen sich insbesondere Schulleitungen als Mittler*innen zwischen Administration und operativer Ebene der Lehrkräfte, wenn es um die Nutzung von Evidenzen, gewonnen aus Instrumenten der Neuen Steuerung, geht (Demski, 2019b).

In Bezug auf VERA lassen sich Einflüsse von Schulleitungen und deren Führungsverhalten auf die Datennutzung der Vergleichsarbeiten ausmachen (z.B. Ercan et al., 2021; Wurster, Bach et

al., 2016). In diesem Kontext untersuchen Kronsfoth, Muslic, Graf und Kuper (2018) den Zusammenhang verschiedener Führungsdimensionen von Schulleitungen und der Nutzung von VERA-Rückmeldungen. Insgesamt finden sich eher wenige signifikante Zusammenhänge zwischen Führungsverhalten und Ergebnisnutzung durch Lehrkräfte. Nur im Hinblick auf eine delegative Führung, also einen Führungsstil, der durch die Übertragung von Aufgaben und Eigenverantwortung an Lehrkräfte gekennzeichnet ist, zeigen sich vereinzelt signifikante negative Korrelationen mit verschiedenen Nutzungsaktivitäten. Darüber hinaus zeigt sich an Schulen mit einer tendenziell direktiv orientierten Schulleitung eine vergleichsweise intensive Rezeption der Ergebnisse durch die Lehrkräfte. Bei einer eher direktiven Führung werden Entscheidungen direkt durch die Schulleitung getroffen und klare Vorschriften gegeben, sodass die Autor*innen die intensivere Rezeption der Rückmeldungen auf eine mögliche Rechenschaftsverpflichtung der Lehrkräfte zu den Ergebnissen der Vergleichsarbeiten zurückführen. Weitere, über die Rezeption hinausgehende Aktivitäten scheinen in keinem Zusammenhang mit einer direktiven Schulleitung zu stehen. Insgesamt gelangen die Autor*innen dennoch zu dem Schluss, dass Schulleitungen durchaus zentrale Akteur*innen für den Anstoß einer datenbasierten Schul- und Unterrichtsentwicklung darstellen können.

Einen anderen Aspekt von Führungsverhalten beleuchten Bach et al. (2014) und ermitteln eine datenbezogene Führungsorientierung (*data-wise leadership)* der Schulleitung, einschließlich der Unterstützung im Umgang mit den rückgemeldeten Daten, als prädiktiv für die Nutzung der VERA-Ergebnisse für Fortbildungsplanungen innerhalb von Fachkonferenzen. Ebenso kommen Kuper et al. (2016) basierend auf Interviewstudien zu dem Schluss, dass Akzeptanz und Nutzung von Vergleichsarbeiten durch entsprechende Maßnahmen der Schulleitung, wie die Bereitstellung angemessener Ressourcen für einen Austausch über die rückgemeldeten Ergebnisse und deren unterrichtsbezogene Verarbeitung, gefördert werden können.

Nicht nur das Verhalten von Schulleitungen beeinflusst den Umgang mit Vergleichsarbeiten, in der Literatur werden verschiedene weitere Kontextfaktoren auf Schulebene aufgezeigt, die berücksichtigt werden sollten. Diese betreffen vor allem die Kooperation im Kollegium sowie schulinterne Kommunikationsstrukturen. Eine ausgeprägte Kooperations- und Evaluationskultur wirken demnach positiv auf den Auseinandersetzungsprozess mit den Rückmeldungen aus Vergleichsarbeiten (Kühle & Peek, 2007; Maier et al., 2011; Wurster, Bach et al., 2016). In einer Untersuchung von Kühle und Peek (2007) zeigt sich an Schulen mit einer fächerübergreifenden Auswertungsstrategie und Nutzung der Rückmeldungen generell ein größeres Interesse an Unterrichtsentwicklung sowie eine stärkere Überzeugung hinsichtlich des Aufschluss-

reichtums, der Bedeutsamkeit und der Nützlichkeit der Ergebnisse aus Vergleichsarbeiten. Auch in anderen Untersuchungen erweisen sich die schulinterne Kommunikation sowie eine kollegiale Datenauswertung bzw. deren Diskussion als förderlich für die Rezeption und weitere Nutzung der Ergebnisse (Maier et al., 2011; Wurster, Bach et al., 2016). Die kooperative Nutzung in Fachkonferenzen ist wiederum mit einem positiven Innovations- bzw. Kooperationsklima assoziiert (Maier et al., 2012).

2.2.2.7. Fazit zum Forschungsstand zur Akzeptanz von VERA

Zusammenfassend lassen sich in der Literatur verschiedenste Wahrnehmungsaspekte identifizieren, die teils auch in den erläuterten Einstellungstheorien und insbesondere im TAM verortet werden können. Neben Untersuchungen zu Akzeptanz, Einstellung und Nützlichkeit (siehe bspw. Diemer, 2013; Maier, 2008a, 2008b; Wagner et al., 2019; Wurster, Bach et al., 2016) ließen sich die Aspekte der zeitlichen Belastung (siehe bspw. Ramsteck & Maier, 2015; Skejic et al., 2015; Vogel, 2020), der Abwägung von Aufwand und Nutzen (siehe bspw. Bonsen et al., 2006; Vogel et al., 2016) sowie weitere Faktoren wie Funktionswahrnehmung der Vergleichsarbeiten und daraus resultierende Konsequenzen (siehe bspw. Jäger, 2011; Richter et al., 2014), Wissen über VERA (siehe bspw. Skejic et al., 2015; Wurster, Bach et al., 2016) oder bestimmte innerschulische Aspekte (siehe bspw. Demski, 2019b; Zlatkin-Troitschanskaia et al., 2016) identifizieren.

Zentrale Erkenntnis der Exzerption der Forschungsliteratur ist, dass sich die identifizierten Wahrnehmungsfaktoren oftmals inhaltlich nicht klar trennen lassen und auch teils wechselseitige Beziehungen zwischen den verschiedenen Aspekten bestehen. Insbesondere die Begriffe Akzeptanz, Einstellung und Nützlichkeit werden hierbei häufig in enger Verknüpfung oder gar synonym verwendet, i. d. R. ohne diese weiter zu definieren oder klar voneinander abzugrenzen. Gerade im Hinblick auf die für das Forschungsinteresse dieser Arbeit relevante Akzeptanz stellt sich heraus, dass keine der zahlreichen Arbeiten, die sich mit der Akzeptanz von VERA beschäftigen, sich darum bemüht, den Akzeptanzbegriff zu klären. Häufig wird das Konstrukt nur über das Zusammenspiel mit anderen Aspekten erklärt, oder die Bedeutung lässt sich, bspw. in quantitativen Arbeiten, aus der Konstruktoperationalisierung ableiten. Generell ist im Hinblick auf die Konstrukte Akzeptanz, Einstellung und Nützlichkeit anzumerken, dass aufgrund unterschiedlicher Operationalisierungen gleich benannter Konstrukte die Ergebnisse verschiedener Studien u. U. nur bedingt vergleichbar sind, da ggf. unter gleichem Namen eigentlich

unterschiedliche Konstrukte gemessen werden. Häufig wird von Akzeptanz gesprochen, inhaltlich werden jedoch Einstellungen oder Nützlichkeitsaspekte gemessen.

Auf Grundlage dieser Erkenntnis und geleitet von der Frage, warum Lehrkräfte Vergleichsarbeiten häufig negativ wahrnehmen und ihnen in der Praxis vielfach ablehnend gegenüberstehen, ist das übergeordnete Ziel dieser Arbeit, eine klare theoriegeleitete Konzeption des Konstrukts Akzeptanz zu entwickeln und diese anhand empirischer Daten zu überprüfen. Der Entwicklung dieser Konzeption einschließlich einer Definition von Akzeptanz widmet sich das folgende Kapitel 3.

3. Ableitung des Forschungsmodells und Hypothesengenerierung

Die Rezeption der Forschungsliteratur zu Vergleichsarbeiten offenbart das Fehlen einer eindeutigen Definition des Akzeptanzbegriffs und es wird klar, dass die Untersuchung von Akzeptanz im Kontext von VERA von konzeptioneller Unschärfe geprägt ist. Wie in Kapitel 2.2.2 ausführlich erläutert, nutzen zwar viele Arbeiten die Bezeichnung Akzeptanz, i. d. R. jedoch ohne diese näher zu definieren, was wiederum zu einer häufigen Vermischung von Begrifflichkeiten und Konstrukten führt. Vielen Arbeiten ist gemein, dass Akzeptanz als zentrale Voraussetzung für eine erfolgreiche Implementierung von Vergleichsarbeiten angesehen wird (siehe bspw. Leutner et al., 2008; Maaz et al., 2019).

Inhaltlich findet der Begriff häufig in Zusammenhang mit Einstellung, Nützlichkeit (siehe bspw. Ditton et al., 2002; Maier, 2008a, 2008b; Wagner et al., 2019) oder sogar Handlungsbereitschaft (Rieß & Zuber, 2014) Verwendung. Ein eindeutiges zugrundeliegendes Begriffsverständnis kann jedoch in vielen Arbeiten nur über die Operationalisierung oder die Ergebnisinterpretation erschlossen werden. Vielfach wird Akzeptanz als ein Einstellungskonstrukt betrachtet, welches i. d. R. die positiven Aspekte von Einstellung abbildet (siehe bspw. Maier, 2008a, 2009c; Wagner et al., 2019). Außerdem besteht in mehreren Untersuchungen eine Verknüpfung mit der Nutzenwahrnehmung, indem Akzeptanz entweder durch Aspekte von Nützlichkeit operationalisiert wird, oder anderweitig eng mit der Nützlichkeitswahrnehmung assoziiert ist (siehe bspw. Maier, 2008b, 2009d).

Die Aufarbeitung der Literatur verdeutlicht, dass sich die Begrifflichkeiten Akzeptanz, Einstellung und Nützlichkeit kaum trennen lassen und es im Kontext der VERA-Forschung viele konzeptionelle Überschneidungen und synonyme Verwendungen gibt. So entwickeln bspw. Ditton et al. (2002) eine Skala zur Messung von Einstellungen von Lehrkräften gegenüber zentralen Tests. Diese Skala, die inhaltlich u. a. Aspekte von Nützlichkeit misst, wird auch in anderen Arbeiten zur Messung von Akzeptanz verwendet (siehe bspw. Groß Ophoff, 2013; Maier, 2009d; Maier et al., 2011; Wagner et al., 2019). Im Gesamten stellt sich ein Zusammenhang von Akzeptanz, Einstellung und Nutzenwahrnehmung heraus, wobei Akzeptanz i. d. R. als ein positives Merkmal im Sinne einer positiven Bewertung oder Einstellung verstanden wird. Wie genau diese Konstrukte zueinander in Beziehung stehen, kann nicht eindeutig geklärt werden.

Einige Arbeiten verstehen Akzeptanz als Teil von Einstellung, andere erachten eine (positive) Einstellung als Bestandteil oder Voraussetzung von Akzeptanz. Allein auf Basis der Forschungsliteratur zu Vergleichsarbeiten war eine Definition von Akzeptanz somit nicht eindeutig möglich.

Aufgrund der hervorgetretenen inhaltlichen Überschneidungen von Akzeptanz und Einstellungen wurde zur weiteren Annäherung an die Thematik aus einer theoretischen Richtung auf Theorien der Einstellungsforschung zurückgegriffen. Einstellungen umfassen demnach affektive, kognitive und verhaltensbezogene Elemente (Eagly & Chaiken, 1993, 1998). Theoretische Basis für das Akzeptanzverständnis in dieser Arbeit bildet das Technology Acceptance Modell (TAM) (siehe Kapitel 2.1.2.4), das diese Elemente in einem Akzeptanzmodell zur Erklärung der Akzeptanz technologischer Innovationen in Zusammenhang bringt. Gemäß dem TAM verhalten sich diese folgendermaßen zueinander: Der Kontakt bzw. die Auseinandersetzung mit einem Akzeptanzobjekt führt zu einer kognitiven Reaktion, stößt also kognitive Evaluationsprozesse an, die zur Herausbildung einer affektiven Einstellungsreaktion führen, auf deren Basis eine behaviorale Reaktion im Sinne einer Nutzungsintention erfolgt. Im TAM fallen die Wahrnehmung von Einfachheit und Nutzen unter kognitive Reaktionen, während Einstellung als affektives Konstrukt verstanden wird, Nutzungsintention und Nutzung beschreiben entsprechend verhaltensbezogene Reaktionen (F. D. Davis, 1993). Aus der Logik des Modells lässt sich die folgende allgemeine Definition von Akzeptanz ableiten: *Akzeptanz beschreibt die Bildung einer positiven (affektiven) Einstellung gegenüber einem Akzeptanzobjekt auf Basis kognitiver Prozesse, die in einer entsprechenden Verhaltensabsicht bzw. einem Verhalten resultiert.*

Akzeptanz ist nach diesem Verständnis somit keine unabhängige Variable, die die Nutzung beeinflusst, und auch keine abhängige Variable, die bspw. von der Nutzenwahrnehmung beeinflusst wird, sondern bildet einen (positiven) Bewertungsprozess ab. Diesem Verständnis nach ist eine Verhaltenskomponente ein inhärenter Bestandteil von Akzeptanz. Akzeptanz endet nicht mit einer positiven Bewertung oder Einstellungsbildung, sondern umfasst immer den nächsten Schritt eines positiven Verhaltensaspekts. Diese Definition und die Struktur der Kausalbeziehungen zwischen den verschiedenen Konstrukten des TAM wurde der Entwicklung eines Akzeptanzmodells zur Erklärung der Akzeptanz von VERA aus Sicht der Lehrkräfte zugrunde gelegt. Dieses Grundmodell wurde mit Hilfe der in Kapitel 2.1.2 gewonnenen Erkenntnisse zur Wahrnehmung von VERA durch Lehrkräfte angepasst und ergänzt. Mit Blick auf die einzelnen Komponenten des TAM lassen sich hierbei viele Parallelen ziehen.

Weitestgehender Konsens herrscht bspw. hinsichtlich des prädiktiven Charakters der Wahrnehmung von Nutzen bzw. Nützlichkeit der Vergleichsarbeiten für verschiedene Aspekte der Ergebnisnutzung (siehe bspw. Demski, 2019b; Groß Ophoff, 2013; Kühle & Peek, 2007; Vogel, 2020; Wurster & Richter, 2016). Zur Beziehung von wahrgenommener Nützlichkeit und Einstellung, wie sie im TAM postuliert wird, gibt es hingegen kaum klare Befunde. Vielfach gibt es Überschneidungen in der Konzeptualisierung, u. a., weil Einstellungen oft mit einer Bewertung von Nützlichkeit gleichgesetzt oder durch Nützlichkeitsaspekte operationalisiert werden (siehe bspw. Vogel et al., 2016; Vogel, 2020; Wurster, Bach et al., 2016). Dies veranschaulicht zwar eine inhaltliche Nähe der Konstrukte im Kontext von VERA, allerdings lassen sich daraus kaum klare Aussagen zu Kausalbeziehungen ableiten. Nur vereinzelt finden sich Hinweise auf einen positiven Zusammenhang zwischen Einstellung und der Einschätzung der Nützlichkeit (Maier, 2008a). Für die Vorhersagefähigkeit von Einstellungen für eine Datennutzung und evidenzbasiertes Handeln finden sich etwas mehr Belege (Demski, 2019b; Muslic, 2017; Vanhoof et al., 2014), Jedoch gibt es auch Erkenntnisse, die belegen, dass eine positive Einstellung nicht zwangsläufig zu einem Anschlusshandeln führen muss (Skejic et al., 2015; Zuber, 2019).

Der weitere für das TAM wichtige kognitive Faktor der wahrgenommenen Einfachheit scheint im Kontext von VERA keine herausragende Rolle zu spielen und wird nur in wenigen Forschungsarbeiten beleuchtet. Möglicherweise liegt dies darin begründet, dass es sich bei den Vergleichsarbeiten um keine technologische Innovation handelt, bei der durch ein mangelndes Verständnis der Anwendung bzw. Durchführung mögliche Hürden einer Nutzung entstehen können. Zwar identifizieren einige Untersuchungen Verständlichkeit von Ergebnisrückmeldungen durchaus als Voraussetzung einer späteren Weiterarbeit (Jäger, 2011; Schneewind & Kuper, 2009) bzw. konstatieren eine Überforderung von Lehrkräften mit den Rückmeldungen von Vergleichsarbeiten (Dedering, 2011; Demski, 2019b; Diemer, 2013), die empirischen Belege hierfür sind jedoch mehrdeutig (Koch, 2011; Kühle & Peek, 2007). Aufgrund des Zeitpunkts der Datenerhebung konnte die Verständlichkeit der Rückmeldungen in dieser Arbeit nicht sinnvoll erhoben werden, da die befragten Lehrkräfte zum Zeitpunkt der Datenerhebung noch keine Gelegenheit zur Rezeption oder gar Reflexion der Ergebnisrückmeldungen hatten, weshalb deren Verständlichkeit noch nicht beurteilt werden kann (siehe zur Erläuterung auch Kapitel 4.1.2). Jedoch zeigt die Analyse der Forschungsliteratur, dass VERA oft als sehr zeitintensiv und dadurch sehr belastend empfunden wird, was einer Weiterarbeit mit den Rückmeldungen entgegensteht (siehe bspw. Diemer, 2013; Ramsteck & Maier, 2015). Vogel (2020) sowie Wurster und Richter (2016) liefern empirische Belege für einen Zusammenhang zwischen

Aufwand und der Wahrnehmung von Nützlichkeit. In diesem Zusammenhang scheint die Abwägung von Aufwand und Nutzen eine besondere Rolle zu spielen. Mehrere Untersuchungen ermitteln ein aus Sicht der beteiligten Akteur*innen ausgeglichenes Verhältnis von Aufwand auf der einen Seite und einem aus der Durchführung resultierenden Nutzen auf der anderen Seite als eine Gelingensbedingung von Vergleichsarbeiten und Voraussetzung für eine Weiterarbeit mit den erhaltenen Daten (siehe bspw. Bonsen et al., 2006; Demski, 2019a, 2019b; Vogel et al., 2016). Aus den aufgeführten Gründen wird die wahrgenommene Einfachheit in dem aufgestellten Modell nicht weiter berücksichtigt und die offensichtlich wichtigen Faktoren zeitliche Belastung und Kosten-Nutzen-Abwägung werden stattdessen in das Modell aufgenommen. Deren Zusammenspiel mit den anderen Modellkonstrukten wird analog zur wahrgenommenen Einfachheit postuliert.

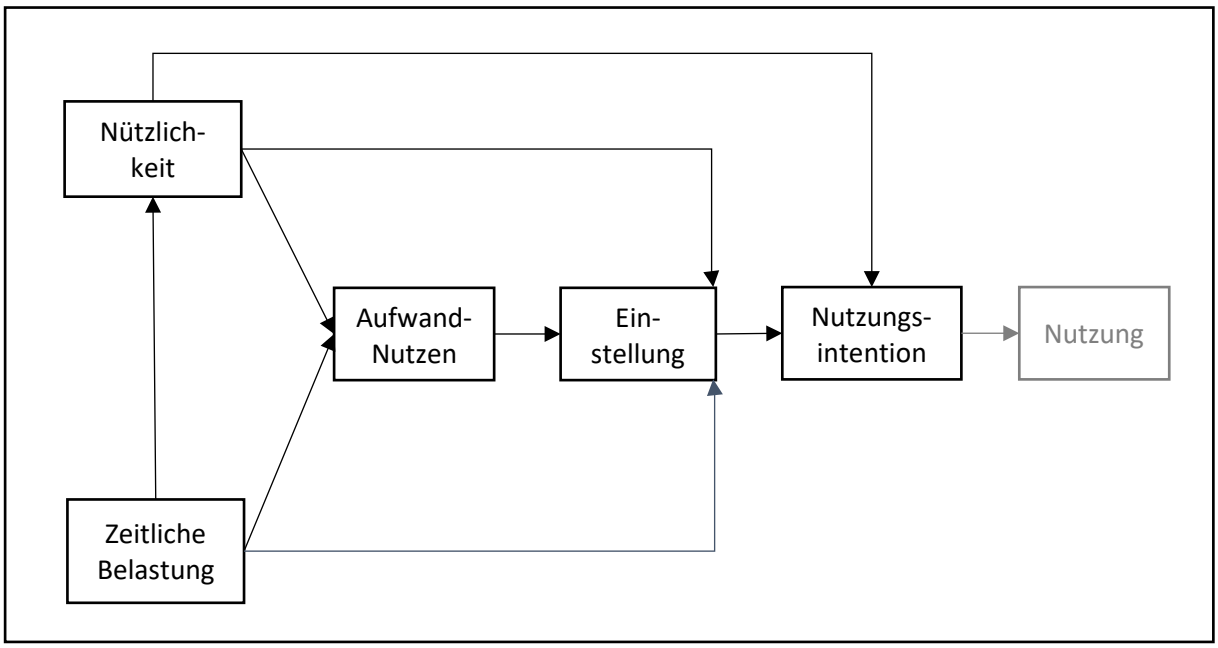


Abbildung 7: Darstellung des entwickelten Forschungsmodells zur Akzeptanz von VERA

Das auf der dargelegten Erkenntnisbasis generierte Forschungsmodell (siehe Abbildung 7), lässt sich folgendermaßen beschreiben: Die zeitliche Belastung wird zum einen analog zur wahrgenommenen Einfachheit im TAM als Prädiktor der wahrgenommenen Nützlichkeit und der Einstellung angesehen. Zusätzlich beeinflusst die wahrgenommene zeitliche Belastung zusammen mit der Wahrnehmung der Nützlichkeit das Aufwand-Nutzen-Verhältnis, welches aufgrund seiner besonderen Relevanz für die Bewertung der Vergleichsarbeiten als separates Konstrukt in das Modell aufgenommen wird. Dieser Teil des Modells spiegelt durch die Abwägung von Aufwand und Ertrag die kognitiven Prozesse bei der Bewertung der Vergleichsarbeiten

wider. Im Rahmen der kognitiven Elemente wird entsprechend den Kausalbeziehungen des TAM zusätzlich eine direkte Wirkung der wahrgenommenen Nützlichkeit auf die Einstellung gegenüber den Vergleichsarbeiten angenommen. Da, in Anlehnung an die Kausalstruktur des TAM, den kognitiven Elementen ein prädiktiver Charakter im Hinblick auf die Einstellung unterstellt wird, wirkt zusätzlich die explizite Aufwand-Nutzen-Abwägung direkt auf die Einstellung. Die Einstellung, welche, entsprechend dem TAM als affektives Konstrukt konzeptualisiert wird, übt im Modell einen direkten Einfluss auf den Nutzungsaspekt, genauer auf die Nutzungsintention, aus. Gemäß dem TAM beeinflusst ausschließlich die Nutzungsintention die tatsächliche Nutzung (F. D. Davis, 1993); eine Annahme, die sich auch in der Metaanalyse von Turner et al. (2010) bestätigt. Die tatsächliche Nutzung konnte in dieser Arbeit aufgrund des Erhebungszeitpunktes nicht erhoben werden (siehe auch Kapitel 4.1.2). Der behaviorale Aspekt von Akzeptanz wird somit durch die Nutzungsintention abgedeckt.

Aus dem aufgestellten Forschungsmodell leiten sich die Hypothesen dieser Arbeit ab. Das postulierte Vorzeichen der in den Hypothesen beschriebenen direkten Pfade ist in Abbildung 8 noch einmal visualisiert.[4]

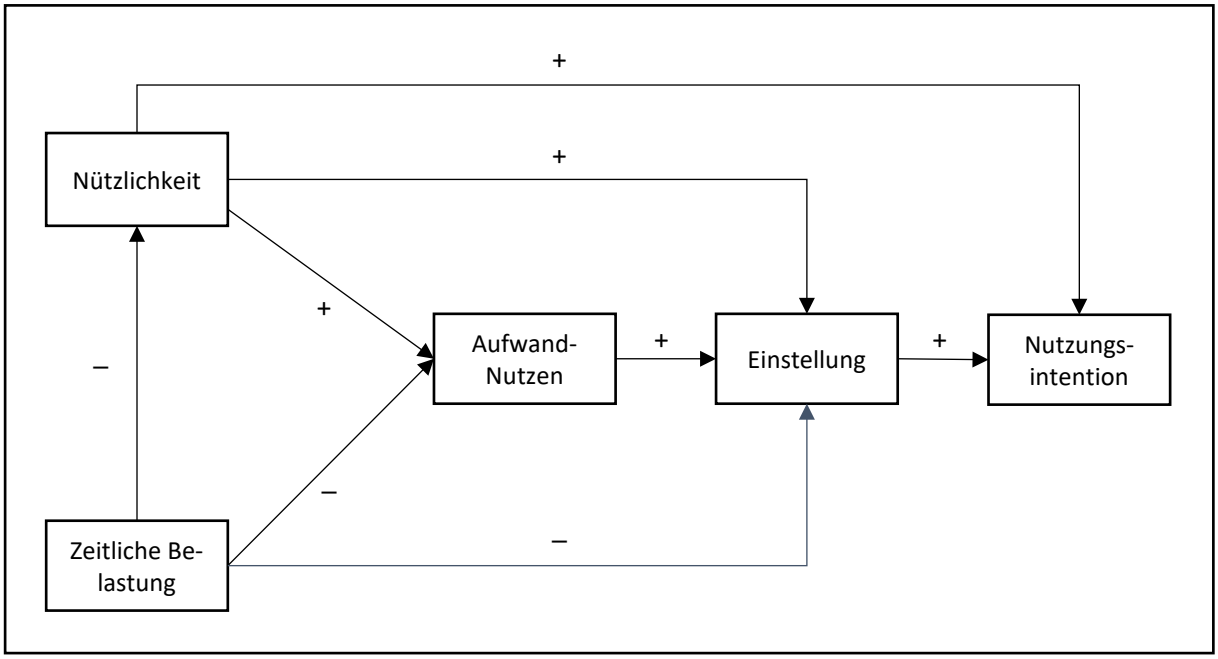


Abbildung 8: Wirkung der Hypothesen

[4] Die Beziehung zwischen Nutzungsintention und Nutzung wurde in dieser Darstellung nicht berücksichtigt, da diese in der vorliegenden Arbeit wie dargelegt nicht untersucht werden konnte.

Die Beschreibung der Hypothesen erfolgt konstruktweise und differenziert zwischen direkten und indirekten Effekten sowie einem Gesamteffekt. Die entsprechenden (Teil-)Hypothesen werden dafür jeweils mit einem d (= direkt), i (= indirekt) bzw. g (= gesamt) gekennzeichnet. Das postulierte Vorzeichen der beschriebenen (Gesamt-)Effekte ist zusätzlich hinter der Hypothesenkurzbeschreibung dargestellt. Gibt es entweder nur einen direkten Effekt oder nur indirekte Effekte, wird die jeweilige Hypothese mit einem d oder i und zusätzlich einem g gekennzeichnet. Dies ist bspw. bei Hypothese H1 der Fall:

H1: Zeitliche Belastung → Wahrgenommene Nützlichkeit (−)

- Je höher die zeitliche Belastung empfunden wird, desto geringer ist die wahrgenommene Nützlichkeit (H1d/g).

H2: Zeitliche Belastung → Aufwand-Nutzen (−)

- Je höher die zeitliche Belastung empfunden wird, desto negativer fällt die Abwägung von Aufwand und Nutzen aus (H2g).
- Hierbei wirkt zum einen die zeitliche Belastung direkt negativ auf die Aufwand-Nutzen-Abwägung (H2d) und zum anderen indirekt, mediiert durch die wahrgenommene Nützlichkeit (H2i).

H3: Zeitliche Belastung → Einstellung (−)

- Je höher die zeitliche Belastung empfunden wird, desto negativer ist die Einstellung der Lehrkräfte (H3g).
- Die zeitliche Belastung wirkt dabei direkt (H3d) und zusätzlich indirekt, mediiert durch die wahrgenommene Nützlichkeit und die Aufwand-Nutzen-Abwägung, negativ auf die Einstellung (H3i).

H4: Zeitliche Belastung → Nutzungsintention (−)

- Die empfundene zeitliche Belastung beeinflusst indirekt, mediiert durch die vorgelagerten Aspekte Nützlichkeit, Aufwand-Nutzen und Einstellung, negativ die Nutzungsintention (H4i/g).

H5: Wahrgenommene Nützlichkeit → Aufwand-Nutzen (+)

- Je ausgeprägter die Wahrnehmung der Nützlichkeit ist, desto positiver fällt die Abwägung von Aufwand und Nutzen aus (H5d/g).

H6: Wahrgenommene Nützlichkeit → Einstellung (+)

- Je ausgeprägter die Wahrnehmung der Nützlichkeit ist, desto positiver ist die Einstellung der Lehrkräfte (H6g).
- Der positive Effekt setzt sich dabei zusammen aus einem direkten positiven Einfluss der wahrgenommenen Nützlichkeit auf die Einstellung (H6d) und einem durch die Wirkung auf die Aufwand-Nutzen-Abwägung mediierten indirekten Effekt (H6i).

H7: Wahrgenommene Nützlichkeit → Nutzungsintention (+)

- Je ausgeprägter die Wahrnehmung der Nützlichkeit ist, desto stärker ist die Nutzungsintention ausgeprägt (H7g).
- Die wahrgenommene Nützlichkeit wirkt dabei zum einen direkt verstärkend auf die Nutzungsintention (H7d) und zum anderen indirekt, mediiert durch die Aufwand-Nutzen-Abwägung und Einstellung (H7i).

H8: Aufwand-Nutzen → Einstellung (+)

- Je positiver die Abwägung von Aufwand und Nutzen ausfällt, desto positiver ist die Einstellung der Lehrkräfte (H8d/g).

H9: Aufwand-Nutzen → Nutzungsintention (+)

- Eine positive Aufwand-Nutzen-Abwägung beeinflusst indirekt, mediiert durch die Einstellung, die Nutzungsintention positiv (H9i/g).

H10: Einstellung → Nutzungsintention (+)

- Je positiver die Einstellung der Lehrkräfte ist, desto stärker ist die Ausprägung der Nutzungsintention der VERA-Rückmeldungen (H10d/g).

Verschiedene Untersuchungen zum TAM belegen einen stärkeren Einfluss der wahrgenommenen Nützlichkeit auf die Nutzungsintention im Vergleich zur Wirkung der wahrgenommenen Einfachheit (Chismar & Wiley-Patton, 2003; King & He, 2006; Ma & Liu, 2004), was zur Ableitung der folgenden Hypothese führt:

H11: Der Gesamteffekt der wahrgenommenen Nützlichkeit auf die Nutzungsintention ist stärker als der Gesamteffekt der zeitlichen Belastung.

Ein weiteres Erkenntnisinteresse (Ziel 4) dieser Arbeit betrifft den Unterschied in der Wahrnehmung und Bewertung der Vergleichsarbeiten durch Lehrkräfte in den verschiedenen Schulformen. Die Aufarbeitung des Forschungsstandes legt offen, dass bisher kaum Arbeiten vorliegen, die die Sicht von Lehrkräften zwischen VERA3 und VERA8 vergleichen. Lediglich Maier und Rauin (2006) beleuchten in ihrer Befragungsstudie mit baden-württembergischen Lehrkräften einige Unterschiede zwischen Grund- und Sekundarschullehrkräften. Grundschullehrkräfte stehen den VERA-Testungen demnach tendenziell aufgeschlossener gegenüber als Lehrkräfte der Sekundarstufe. Derartige Unterschiede spiegeln auch Erfahrungswerte aus der Projektpraxis der vom Zentrum für Empirische Pädagogische Forschung (zepf) durchgeführten VERA-Durchgänge wider. Diese legen nahe, dass zwischen Grundschullehrkräften und Lehrkräften weiterführender Schulen durchaus Unterschiede in der Bewertung der Vergleichsarbeiten bestehen. Dies zeigt sich sowohl in den Rückmeldungen bzw. Supportanfragen von Lehrkräften zu den VERA-Durchführungen als auch im Antwortverhalten in bisher unveröffentlichten Evaluationsbefragungen des zepf mit VERA3- und VERA8-Lehrkräften, bspw. aus dem Jahr 2017 (siehe Kapitel 4.1 für die Beschreibung vergleichbarer Befragungen). In der deskriptiven Auswertung zeigt sich, dass Grundschullehrkräfte Fragen zur Bewertung der Vergleichsarbeiten im Schnitt positiver beantworten als Lehrkräfte, die an VERA8 teilgenommen haben und auch tendenziell eine geringere zeitliche Belastung im Zuge der Durchführung von Vergleichsarbeiten empfinden. Auf Basis dieser Erfahrungen der VERA-Praxis soll in dieser Arbeit der Frage nach Unterschieden in der Bewertung der Vergleichsarbeiten zwischen den Schularten empirisch nachgegangen und potenzielle Unterschiede sollen im Detail analysiert werden. Auf die Formulierung expliziter Hypothesen wird jedoch an dieser Stelle aufgrund einer fehlenden empirischen und theoretisch belastbaren Grundlage verzichtet, sodass dieser Teil der Untersuchung (siehe Kapitel 5.4) einen eher explorativen Charakter aufweist.

4. Methodisches Vorgehen

In diesem Kapitel wird das methodische Vorgehen der empirischen Untersuchung vorgestellt. Im Rahmen der Beschreibung des Untersuchungsdesigns (Kapitel 4.1) wird zunächst die Operationalisierung der untersuchten Konstrukte dargelegt (Kapitel 4.1.1), bevor das Vorgehen der Datenerhebung (Kapitel 4.1.2) erläutert wird. Kapitel 4.1.3 umreißt das Vorgehen bei der Auswertung der Daten zur Beantwortung der Forschungsfragen und Hypothesen. In Kapitel 4.2 erfolgt eine ausführliche Beschreibung der Stichprobe einschließlich der Aufbereitung der Daten. Kapitel 4.2.1 beschäftigt sich mit den Besonderheiten von Survey-Studien im Hinblick auf das Auftreten fehlender Werte. Kapitel 4.2.2 erläutert die Rücklaufquote der Datenerhebung, die Teilkapitel 4.2.3 und 4.2.4 die Datensatzaufbereitung und in Kapitel 4.2.5 erfolgt die Beschreibung der finalen Stichprobe. Kapitel 4.3 beschäftigt sich schließlich mit den verwendeten Analysemethoden. Der Aufbau dieses Kapitels orientiert sich hierbei an den Aspekten, die für ein Verständnis der statistischen Analysen in dieser Arbeit notwendig sind, mit Fokus auf entsprechenden komplexeren multivariaten Verfahren und der Darstellung entsprechender Kennzahlen. Kapitel 4.3.1 begründet zunächst die Auswahl der gewählten Analysemethode einer Strukturgleichungsmodellierung. Diese wird in den Kapiteln 4.3.2 und 4.3.3 ausführlich dargestellt. Das letzte Teilkapitel 4.3.4 widmet sich der Erläuterung latenter Gruppenmodelle.

4.1. Untersuchungsdesign

4.1.1. Erhebungsinstrument: Operationalisierung der Modellkonstrukte

Um die untersuchten Konstrukte angemessen erfassen zu können, wurde zunächst ein entsprechendes Messinstrument entwickelt, welches im Folgenden beschrieben ist. Da es sich bei derartigen Konstrukten typischerweise um eine Art abstraktes Konzept handelt, das nicht direkt beobachtet bzw. gemessen werden kann, ist dieser Schritt der Operationalisierung notwendig. Hierzu werden das Konstrukt beschreibende bzw. widerspiegelnde Vorgänge oder Tätigkeiten definiert, die im Gegensatz zu dem Konstrukt selbst, direkt beobachtbar bzw. messbar sind (Gay et al., 2011).

Für diese Untersuchung wurde im Zuge der Operationalisierung ein weitestgehend neues Erhebungsinstrument entwickelt. Ziel war es dabei, in Anlehnung an das TAM und unter Anpassung

an die Anforderungen des Untersuchungsgegenstandes, die verschiedenen Aspekte der Wahrnehmung von Vergleichsarbeiten durch Lehrkräfte adäquat zu erfassen. Zu diesem Zweck wurden die in Kapitel 2.2.2 erarbeiteten Erkenntnisse zur Wahrnehmung von Vergleichsarbeiten durch Lehrkräfte sowie Items und Skalen aus bisher teils unveröffentlichten Lehrkräftebefragungen des zepf herangezogen. Aufgabe des Instruments ist es, die folgenden Konstrukte in Bezug auf Vergleichsarbeiten zu erfassen: wahrgenommene Nützlichkeit, wahrgenommene zeitliche Belastung, Aufwand-Nutzen-Verhältnis, Einstellung sowie die Nutzungsintention der Lehrkräfte.

Nützlichkeit

Das Konstrukt Nützlichkeit erfasst einen der kognitiven Aspekte im Bewertungsprozess der Vergleichsarbeiten. Inhaltlich bezieht sich das Konstrukt auf verschiedene Gesichtspunkte der praktischen Nützlichkeit der durch die Vergleichsarbeiten erhaltenen Daten, die auch im Literaturreview (siehe Kapitel 2.2.2) herausgearbeitet wurden. Bei den Items, die bereits seit 2017 in den Evaluationsbefragungen des zepf verwendet werden, handelt es sich um Eigenkonstruktionen. Insgesamt wird das Konstrukt Nützlichkeit mit fünf Items abgebildet, die jeweils auf einer 4-stufigen Likert-Skala von 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu* gemessen werden (siehe Tabelle 1).

Tabelle 1: Operationalisierung des Konstrukts Nützlichkeit

Abkürzung des Konstrukts	Item	Itemtext
WN	WN1	Die rückgemeldeten Daten… haben einen praktischen Nutzen für meinen Unterricht.
	WN2	Die rückgemeldeten Daten… liefern mir zusätzliche Informationen, die ich durch eigene Diagnose im Unterricht nicht erhalte.
	WN3	Die rückgemeldeten Daten… dienen mir als Ansatzpunkte zur Planung individueller Fördermaßnahmen.
	WN4	Die rückgemeldeten Daten… bieten Anregungen zum pädagogischen Austausch mit Kolleginnen und Kollegen.
	WN5	Die rückgemeldeten Daten… sind nützlich für die Einschätzung der Leistungen einzelner Schüler*innen.

Anmerkung. Skala: 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu.*

Eine gut konstruierte Likert-Skala kann trotz ihres eigentlich ordinalen Charakters wie eine Intervallskala behandelt werden und eignet sich somit auch für die entsprechenden statistischen Verfahren wie bspw. die in dieser Arbeit genutzten Strukturgleichungsmodelle. Bei der Konstruktion einer Likert-Skala sollte daher auf folgende Kriterien geachtet werden: die Skala sollte als symmetrisch wahrgenommen werden, d. h. es gibt ebenso viele positive wie negative Kategorien, die ggf. bei einer ungeraden Anzahl an Kategorien um eine neutrale Bewertungskategorie angeordnet sind. Des Weiteren wird eine identische Distanz zwischen den einzelnen nebeneinander liegenden Ausprägungen gefordert sowie eine sprachliche Klarheit in der Unterscheidung der Kategorien (Hair et al., 2017). Die in dieser Arbeit verwendeten Skalen entsprechen diesen Anforderungen.

Zeitliche Belastung

Einen weiteren kognitiven Aspekt in der Beurteilung von VERA beschreibt die damit einhergehende zeitliche Belastung (siehe Tabelle 2).

Tabelle 2: Operationalisierung des Konstrukts zeitliche Belastung

Abkürzung des Konstrukts	Item	Itemtext
ZB	ZB1	Wie bewerten Sie den Zeitaufwand für die Vorbereitung?
	ZB2	Wie bewerten Sie den Zeitaufwand für die Durchführung?
	ZB3	Wie bewerten Sie den Zeitaufwand für die Auswertung?

Anmerkung. Skala: 1 = *sehr gering* bis 5 = *sehr hoch*.

Das Konstrukt repräsentiert die im ursprünglichen TAM untersuchte wahrgenommene Einfachheit. Entgegen des in den meisten TAM Untersuchungen positiv gepolten Konstrukts wahrgenommene Einfachheit wurde in dieser Arbeit das Konstrukt zeitliche Belastung mit negativ gepolten Indikatorvariablen erfasst. Ziel war es, den in der Literaturaufarbeitung hervorgetretenen Kritikpunkt einer außerordentlichen Zusatzbelastung durch die Durchführung der Vergleichsarbeiten, der von vielen Lehrkräften vorgebracht wird, abzubilden. Das Konstrukt wurde mit 3 Items gemessen, die den Aufwand der einzelnen Durchführungsschritte erfassen. Bei der Skala handelt es sich dabei um eine 5-stufige Likert-Skala von 1 = *sehr gering* bis 5 = *sehr hoch*.

Aufwand-Nutzen

Das Konstrukt Aufwand-Nutzen (AN) spiegelt die im Hinblick auf VERA häufig stattfindende Abwägung von zeitlichem Mehraufwand und empfundenem Nutzen für den eigenen Unterricht wider. Zur Messung des Konstrukts wurden auf Basis der Erkenntnisse verschiedener Forschungsarbeiten (z.B. Bonsen et al., 2006; Demski, 2019a, 2019b; Muslic, 2017) insgesamt sechs Items generiert (siehe Tabelle 3).

Tabelle 3: Operationalisierung des Konstrukts Aufwand-Nutzen

Abkürzung des Konstrukts	Item	Itemtext
AN	AN1 [a]	Betrachte ich den Nutzen, den ich aus den Vergleichsarbeiten für meinen Unterricht ziehe, empfinde ich VERA als zu zeitintensiv.
	AN2 [a]	Die Vergleichsarbeiten kosten Zeit, sind aber nützlich.
	AN3 [a]	Die Zeit, die ich für die Durchführung und Auswertung von VERA sowie für die Auseinandersetzung mit den Ergebnissen benötige, ist sinnvoll investiert.
	AN4 [b]	Im Verhältnis zu den gewonnenen Erkenntnissen empfinde ich den Zeitaufwand der…Vorbereitung von VERA als…
	AN5 [b]	Im Verhältnis zu den gewonnenen Erkenntnissen empfinde ich den Zeitaufwand der…Durchführung von VERA als…
	AN6 [b]	Im Verhältnis zu den gewonnenen Erkenntnissen empfinde ich den Zeitaufwand der…Auswertung von VERA als…

Anmerkungen. [a] Skala: 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu;* [b] Skala: 1 = *nicht angemessen* bis 4 = *angemessen.*

Die Items AN1 bis AN3 wurden jeweils mit einer 4-stufigen Likert-Skala von 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu* gemessen, wobei das Item AN1 negativ gepolt ist und für die spätere Auswertung umgepolt wurde. Die Items AN4 bis AN6 wurden auf einer Likert-Skala mit 1 = *nicht angemessen* bis 4 = *angemessen* erfasst.

Einstellung

Die Einstellung (AE) wurde in dieser Arbeit als ein affektiv konnotiertes Konstrukt im Sinne einer generellen eher intuitiv geprägten Zustimmung oder Ablehnung der Vergleichsarbeiten verstanden und entsprechend operationalisiert (siehe Tabelle 4). Der Itemgruppe zur Einstellungsmessung ging die folgende Anweisung voran: „Bitte bewerten Sie VERA ganz intuitiv gemäß Ihrer persönlichen Einschätzung und entscheiden Sie aus dem Bauch heraus." Gemessen wurde die Einstellung mit insgesamt fünf Items. Vier davon (Item AE1 bis AE4) wurden, anders als die übrigen Konstrukte in dieser Arbeit, mit Hilfe eines 4-stufigen semantischen Differenzials erfasst. Bei dieser Skala handelt es sich um ein typisches Instrument zur Einstellungsmessung. Dieser Skalentyp erfordert eine Einschätzung zu bestimmten Eigenschaften eines bestimmten Themas, Konstrukts etc. auf einem Kontinuum bipolarer Adjektivpaare, wobei jede Position des Kontinuums, wie bei einer Likert-Skala, mit einem bestimmten Wert assoziiert ist (Gay, Mills & Airasian, 2011). Zusätzlich wurde mit einem weiteren Item die generelle Zustimmung zur Durchführung der Vergleichsarbeiten, wieder mit einer 4-stufigen Likert-Skala, erhoben (Item AE5).

Tabelle 4: Operationalisierung des Konstrukts Einstellung

Abkürzung des Konstrukts	Item	Itemtext
AE	AE1 [a]	Ich empfinde die Vergleichsarbeiten als ... nutzlos – nützlich
	AE2 [a]	Ich empfinde die Vergleichsarbeiten als ... negativ – positiv
	AE3 [a]	Ich empfinde die Vergleichsarbeiten als ... nicht sinnvoll – sinnvoll
	AE4 [a]	Ich empfinde die Vergleichsarbeiten als ... nicht hilfreich – hilfreich
	AE5 [b]	Ich befürworte die Durchführung der Vergleichsarbeiten.

Anmerkungen. [a] Skala: semantisches Differenzial von 1 bis 4; [b] Skala: 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu*.

Nutzungsintention

Aufgrund des frühen Erhebungszeitpunktes konnte der Nutzungsaspekt, wie bereits in Kapitel 3 beschrieben, nur durch die geplante Nutzung, also die Nutzungsintention (NI) erfasst werden. Das Konstrukt Nutzungsintention wurde durch den Selbstbericht geplanter Aktivitäten im

Hinblick auf die VERA-Rückmeldungen operationalisiert (siehe Tabelle 5). Die operationalisierten Aktivitäten stellen dabei zentrale Bestandteile von Unterrichtsentwicklung auf Basis von VERA dar, die im Literaturreview in Kapitel 2.2.2 herausgearbeitet wurden. In Anlehnung an das Prozessmodell von Helmke und Hosenfeld (2005) zielt das Konstrukt mit Item NI4 sowohl auf Aspekte der Rezeption und Reflexion ab als auch durch die übrigen Items auf die Aktion in Folge von VERA. Einige der Items wurden auch bereits so oder in abgewandelter Form in früheren Evaluationsbefragungen des zepf genutzt. Alle Items sind daher Eigenkonstruktionen, die jedoch auf Erkenntnissen zur Ergebnisnutzung von Vergleichsarbeiten und früheren Befragungen basieren. Insgesamt wurde das Konstrukt Nutzungsintention mit fünf Items erfasst, die jeweils auf einer 4-stufigen Likert-Skala gemessen wurden.

Tabelle 5: Operationalisierung des Konstrukts Nutzungsintention

Abkürzung des Konstrukts	Item	Itemtext
NI	NI1	Im Hinblick auf die Vergleichsarbeiten plane ich in diesem Jahr ... die VERA-Ergebnisse für meine Unterrichtsplanung zu nutzen.
	NI2	Im Hinblick auf die Vergleichsarbeiten plane ich in diesem Jahr ... basierend auf den VERA-Ergebnissen mit meinen Schüler*innen (verstärkt) auf die in den Bildungsstandards beschriebenen Kompetenzen hinzuarbeiten.
	NI3	Im Hinblick auf die Vergleichsarbeiten plane ich in diesem Jahr ... aus den VERA-Ergebnissen konkrete Veränderungen für meinen Unterricht abzuleiten.
	NI4	Im Hinblick auf die Vergleichsarbeiten plane ich in diesem Jahr ... mich (intensiv) mit den VERA-Ergebnissen auseinander zu setzen.
	NI5	Im Hinblick auf die Vergleichsarbeiten plane ich in diesem Jahr ... bestimmte Aufgabentypen für meinen Unterricht zu übernehmen.

Anmerkung. Skala: 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu*.

4.1.2. Erhebungsmethode und Durchführung der Datenerhebung

Die im vorangegangen Unterkapitel vorgestellten Konstrukte wurden in einen umfangreicheren Fragebogen eingebettet, dessen übrige Fragen jedoch für diese Arbeit nicht relevant sind. Die

Datenerhebung mit diesem Fragebogen ist im Folgenden beschrieben: Die Datengrundlage zur Beantwortung der empirischen Fragestellung bilden Daten aus Lehrkräftebefragungen, die im Kontext der vom Zentrum für Empirische Pädagogische Forschung (zepf) an der Rheinland-Pfälzisch Technischen Universität Kaiserslautern-Landau (RPTU) administrierten Durchführung der Vergleichsarbeiten erhoben wurden. In sogenannten Evaluationsbefragungen wurden Lehrkräfte in der Primarstufe nach der Teilnahme an VERA3 in den Jahren 2018 und 2019 befragt sowie Lehrkräfte in der Sekundarstufe I, die 2018 mit mindestens einer Lerngruppe am VERA8-Durchgang dieses Jahres teilgenommen hatten. Die Befragung erfolgte online und wurde mit Hilfe einer lizensierten Version der Online-Umfrage-Applikation Limesurvey in der Version 2.06+ (siehe Limesurvey GmbH) umgesetzt und durchgeführt. Die Lehrkräfte wurden nach der Durchführung der Vergleichsarbeiten bzw. direkt nach der Eingabe der Ergebnisse gebeten, online einen Fragebogen auszufüllen, der Ihnen über einen Link im VERA-Portal des zepf zur Verfügung gestellt wurde. Die Teilnahme an der Evaluationsbefragung erfolgte vollständig anonym und freiwillig.

Die Befragungen zu VERA3 wurden sowohl 2018 als auch 2019 jeweils in den Bundesländern Bremen, Niedersachsen, Rheinland-Pfalz, Saarland, Schleswig-Holstein, Nordrhein-Westfalen, Mecklenburg-Vorpommern sowie bei der Deutschsprachigen Gemeinschaft in Belgien durchgeführt. Hierbei gilt zu beachten, dass die Verpflichtung zur Teilnahme an den Vergleichsarbeiten im Jahr 2019 in Niedersachsen aufgehoben wurde (siehe auch Kapitel 2.2.1), was sich deutlich auf die Anzahl der teilnehmenden Lerngruppen an den Vergleichsarbeiten und somit auch auf die Zahl der Teilnehmenden bei der Evaluationsbefragung auswirkte. In der Sekundarstufe I wurden 2018 Lehrkräfte aus Bremen, Niedersachsen, Rheinland-Pfalz und dem Saarland zu VERA8 befragt.

Tabelle 6: Erhebungszeiträume

VERA-Durchführung	Zeitraum der Befragung
VERA8 2018	20. Februar – 6. Juni 2018
VERA3 2018	17. April – 3. Juni 2018
VERA3 2019	24. April – 23. Juni 2019

Der Zeitraum der Erhebung erstreckte sich jeweils vom Beginn des jeweiligen Dateneingabezeitraums im Onlineportal direkt nach der VERA-Durchführung bis zum Ende des

Rückmeldezeitraums, nachdem keine weiteren Besuche des VERA-Portals durch Lehrkräfte und somit keine weiteren Teilnahmen mehr zu erwarten waren. Die genauen Erhebungszeiträume sind in Tabelle 6 nachzulesen.

In den vorangegangenen Jahren hatte sich gezeigt, dass die Bereitschaft der Lehrkräfte, an einer sich den Vergleichsarbeiten anschließenden Evaluationsbefragung teilzunehmen, deutlich zurückgegangen war. Im Vergleich zu den Stichproben in den Arbeiten von Groß Ophoff (2013) und Koch (2011), die mit Daten aus den Jahren 2004 bis 2008 arbeiten, die, ebenso wie diese Arbeit, die jährlich vom zepf durchgeführten Evaluationsbefragungen als Datengrundlage nutzen, nahmen die Zahlen der aufgerufenen bzw. ausgefüllten Befragungen bis zum Jahr 2015 sowohl bei den VERA3- als auch bei den VERA8-Lehrkräften deutlich ab. Basieren die Arbeiten von Groß Ophoff (2013) und Koch (2011) noch auf Stichproben von rund 1200 bis teils sogar 3000 ausgefüllten Fragebögen, lagen die Zahlen in den darauffolgenden Jahren zunehmend darunter. Zudem konnte bereits Groß Ophoff (2013) in ihrer Arbeit eine über die Jahre abnehmende Bereitschaft zur Teilnahme an einer VERA-Evaluationsbefragung feststellen.

Diese früheren Befragungen wurden erst mit einigem Abstand zur VERA-Durchführung nach Erhalt der Rückmeldungen angeboten, was für die Lehrkräfte ein erneutes Anmelden im VERA-Portal erforderte. Aufgrund der abnehmenden Teilnahmebereitschaft und um wieder eine größere Stichprobe zu erzielen, wurde der Durchführungszeitraum der Evaluationsbefragungen auf das beschriebene Zeitfenster vorgezogen, um eine Teilnahmehürde zu reduzieren, da das Abrufen der VERA-Ergebnisse ohnehin eine Anmeldung im Portal erfordert.

Der frühe Erhebungszeitpunkt direkt nach der Durchführung bzw. Dateneingabe hat jedoch auch einige Nachteile. Da die Befragung i. d. R. nun vor Erhalt der Rückmeldungen erfolgte, konnten nicht alle Fragen, wie bspw. zu den Rückmeldungen, ohne weiteres beantwortet werden, wenn noch keine Vorerfahrungen aus den vorangegangen VERA-Durchgängen bestanden. Daher wurden die Lehrkräfte in der Befragung bei den betroffenen Itemgruppen explizit aufgefordert auch auf indirekte Erfahrungen, bspw. durch Erfahrungsberichte von Kolleg*innen oder den Austausch in Fachgruppen, zurückzugreifen. Hierfür wurde an den entsprechenden Stellen im Fragebogen der folgende Hinweis integriert: „Bevor Sie die folgenden Fragen beantworten, erinnern Sie sich bitte an die VERA-Durchgänge in den letzten Jahren und beziehen Sie zur Beantwortung alle Erfahrungen (inklusive Erfahrungen innerhalb der Fachgruppe bzw. Schule), die Sie bisher mit VERA gemacht haben, mit ein.“. Zusätzlich hatten die Lehrkräfte, wie bei

allen Items, die Möglichkeit, die Fragen nicht zu beantworten bzw. bei diesen Items die Antwortoption „kann ich nicht beurteilen“ auszuwählen.

4.1.3. Auswertungsdesign

Das folgende Kapitel soll zur Orientierung für den weiteren Verlauf dieser Arbeit dienen und widmet sich der Beschreibung der Analyseschritte zur Beantwortung der aufgestellten Forschungsfragen und zur Überprüfung der in Kapitel 3 im Rahmen des Forschungsmodells formulierten Hypothesen. Hierbei erfolgt eine grobe Skizzierung der durchgeführten Analysen einschließlich einer kurzen Erläuterung der für die jeweiligen Schritte genutzten Datenbasis. Ausführlich werden die Stichproben und Auswertungsmethoden in den Kapiteln 4.2.5 und 4.3 behandelt.

Zunächst erfolgte die empirische Überprüfung des in Kapitel 3 aufgestellten ursprünglichen Forschungsmodells mithilfe von Strukturgleichungsmodellierungen (siehe Kapitel 5.1). Das Vorgehen bei dieser statistischen Methode wird in Kapitel 4.3 ausführlich erläutert. Die Datengrundlage für diese Analysen bildet die VERA3-Evaluationsbefragung des Jahres 2018 (siehe Kapitel 4.2.5 Beschreibung der Stichprobe: Stichproben der Modelltestung und Validierung). Im nächsten Schritt erfolgte eine, in Kapitel 5.2 beschriebene Revision des Ursprungsmodells. Die spezifizierten Anpassungen wurden zunächst mit derselben Datengrundlage wie das ursprüngliche Modell überprüft. Dieses angepasste Modell wurde, wie in Kapitel 5.3 berichtet, anhand eines unabhängigen Datensatzes validiert. Hierfür wurden Daten aus der Evaluationsbefragung im Zuge der VERA3-Durchführung des Jahres 2019 genutzt (siehe ebenfalls Kapitel 4.2.5).

Zur Beantwortung der Forschungsfrage nach Unterschieden in der Akzeptanz zwischen VERA3- und VERA8-Lehrkräften wurde in Kapitel 5.4 eine latente Gruppenanalyse gerechnet. Das Vorgehen bei dieser statistischen Methode wird in Kapitel 4.3.4 beschrieben. Die Datengrundlage für die Schätzung eines Gruppenmodells bilden wiederum Daten aus den Evaluationsbefragungen des Jahres 2018. Neben den bereits für die ursprüngliche Modellschätzung und Modellanpassung verwendeten VERA3-Evaluationsdaten, bzw. einer daraus extrahierten Teilstichprobe, wurden Daten aus der Evaluationsbefragung im Zuge der VERA8-Testung genutzt. Im Detail wird die Stichprobe in Kapitel 4.2.5 (Beschreibung der Stichprobe: Stichproben der latenten Zweigruppenanalyse) beschrieben.

4.2. Datensatzbereinigung und Stichprobe

4.2.1. Fehlende Werte bei Survey-Studien

Die Durchführung von Survey-Studien, insbesondere Onlinebefragungen, birgt einige Besonderheiten, die vor allem das Fehlen von Daten betreffen und somit Auswirkungen auf die in einer solchen Befragung gewonnenen Stichprobe haben können. Hierbei wird zwischen dem Fehlen ganzer Analyseeinheiten (*Unit-Nonresponse*) und dem Fehlen einzelner Daten bei einem vorliegenden Fall (*Item-Nonresponse*) unterschieden.

Unit-Nonresponse

Bei Unit-Nonresponse kommt es durch Nicht-Beobachten von Personen oder Personengruppen zum Ausfall ganzer Analyseeinheiten. Unit-Nonresponse bezieht sich auf die Teilnahmequote, die angibt, wie viele der ursprünglich angesprochenen Personen an einer Umfrage teilgenommen haben. Die Hauptgründe dieser Komplettausfälle liegen in der Nichterreichbarkeit potenzieller Teilnehmender oder in der Verweigerung einer Teilnahme (Engel & Schmidt, 2014). Hinsichtlich der Nichterreichbarkeit spielen bei Onlinebefragungen insbesondere *Abdeckungs-* und *Stichprobenfehler* (*Coverage Error* bzw. *Sampling Bias*) eine Rolle. Ein Abdeckungsfehler entsteht, wenn die Zielpopulation, auf die sich das Forschungsinteresse richtet, von der Erhebungsgesamtheit abweicht. Dies wäre bspw. bei einer Kontaktierung der Zielpopulation per E-Mail der Fall, wenn nicht sichergestellt ist, dass alle potenziellen Proband*innen eine Mailadresse besitzen und somit auch erreicht werden können. Stichprobenfehler entstehen typischerweise, wenn anstatt der Grundgesamtheit eine evtl. verzerrte Stichprobe untersucht wird, weil nicht alle potenziell Teilnehmenden Zugang zur Umfrage haben (Couper, 2000; Kaczmirek, 2008; Umbach, 2004).

Durch die Verweigerung einer Teilnahme entsteht die Gefahr des Auftretens eines *Nonresponse-Error*, wenn sich das Antwortverhalten von Teilnehmenden systematisch vom hypothetischen Antwortverhalten von Nicht-Teilnehmenden unterscheidet (Kaczmirek, 2008). Allein dadurch, dass sie auf die Aufforderung an einer Befragung teilzunehmen nicht reagieren, unterscheiden sich Nicht-Teilnehmende von den Proband*innen. Möglicherweise weist eine der Gruppen ein größeres oder schwächeres Interesse am betreffenden Thema auf (Gay et al., 2011). Generalisierbare Aussagen können daher nur getroffen werden, wenn sich die

Teilnehmenden in ihren Charakteristika nicht signifikant von den Nicht-Teilnehmenden unterscheiden (Tuten, Urban & Bosnjak, 2002).

Item-Nonresponse

Ein weiteres Problem von Onlinebefragungen betrifft die Vollständigkeit der letztendlich vorliegenden Daten. Ein typisches Charakteristikum von Befragungsstudien ist das Vorliegen unvollständiger Datensätze durch Item-Nonresponse, also das nur lückenhafte Ausfüllen durch einzelne Teilnehmende (Schafer & Graham, 2002). Möglicherweise will oder kann ein*e Proband*in aus verschiedenen Gründen bestimmte Fragen nicht beantworten oder ermüdet gegen Ende des Fragebogens und bricht die Befragung ab. Hierdurch ergeben sich verschiedene Muster unvollständig ausgefüllter Fragebögen und entsprechender fehlender Werte in den Datensätzen (Engel & Schmidt, 2014; Graham, 2012). Item-Nonresponse könnte zwar in Onlinebefragungen durch eine in der Befragungssoftware integrierte Verpflichtung zum Ausfüllen aller Items entgegengewirkt werden, ein vorzeitiges Abbrechen der Befragung würde dadurch jedoch nicht verhindert bzw. dieses Verhalten dadurch sogar tendenziell gefördert (Nayak & Narayan, 2019). Daher wurde in den für diese Arbeit relevanten Erhebungen auf eine verpflichtende Beantwortung aller Items verzichtet.

Die American Association for Public Opinion Research (AAPOR, 2016) unterscheidet bei der Bewertung des Anteils fehlender Werte zwischen vollständiger Bearbeitung, partieller Bearbeitung und dem vorzeitigen Abbruch einer Befragung. Bei einer vollständigen Bearbeitung werden gemäß dieser Definition alle Fragen, d. h. alle zur Beantwortung der zu untersuchenden Forschungsfragen als relevant und unerlässlich definierten Fragen, beantwortet. Relevante Items sind dabei bspw. diejenigen, die als entscheidende abhängige oder unabhängige Variablen in einer Studie untersucht werden. Wurden Minimum 50 % der relevanten Items beantwortet, gilt der Fragebogen als hinreichend ausgefüllt. Eine Bearbeitung von unter 50 % der Items gilt hingegen als nicht ausreichend, sodass die entsprechenden Fälle aus dem Datensatz entfernt werden sollten. Die Anwendung dieser Heuristik auf die Datengrundlage dieser Arbeit ist in Kapitel 4.2.4 beschrieben. Unvollständige Fälle, die nach der Aufbereitung in den Daten verblieben, wurden in der statistischen Auswertung im Rahmen der Strukturgleichungsanalyse entsprechend berücksichtigt (siehe Kapitel 4.3.3).

4.2.2. Rücklaufquote und Repräsentativität

Rücklaufquote

Ein häufiges Problem von Online-Befragungen ist die Teilnahmequote, die typischerweise deutlich geringer ausfällt als bei einer Paper-Pencil-Befragung (Nayak & Narayan, 2019). Die bereits angesprochene Problematik der Erreichbarkeit potenzieller Teilnehmenden durch Coverage Error sowie Sampling Bias sollte dabei in dieser Arbeit kaum ein Problem darstellen: Durch die Platzierung des Fragebogens im VERA-Portal, in welchem auch die Eingabe der Testergebnisse erfolgt, wurden alle Lehrkräfte, die im jeweiligen Jahr an VERA teilgenommen haben, gleichermaßen angesprochen und zur Teilnahme aufgefordert. Da so allen potenziellen Teilnehmenden der gleiche Zugriff gewährt wurde, stellt eine Nichterreichbarkeit vermutlich eine vernachlässigbare Ursache für Nichtteilnahme dar. In einigen Fällen gilt jedoch zu berücksichtigen, dass nicht alle Lehrkräfte die Dateneingabe selbst vornehmen, sondern in einigen Schulen diese Aufgabe bspw. vom Sekretariat übernommen wird, was sich negativ auf die Teilnahmequote auswirken könnte. Hierfür liegen jedoch keine Daten vor, die eine Abschätzung der Bedeutsamkeit dieses Anteils ermöglichen würden.

Durch die vollständig anonyme Erhebung der Daten kann die Teilnahmequote der einzelnen Befragungen nur geschätzt werden. Die tatsächliche Grundgesamtheit der Untersuchung stellen diejenigen Lehrkräfte dar, die im Jahr der jeweiligen Erhebung mit mindestens einer Lerngruppe an VERA teilgenommen hatten. Da hierüber keine genauen Informationen vorliegen, wurde anstelle der Anzahl der Lehrkräfte die Anzahl der teilnehmenden Lerngruppen, als Näherung herangezogen. Allerdings kann bei VERA8, nicht ausgeschlossen werden, dass eine Lehrkraft mehreren Lerngruppen zuzuordnen ist, da dort nicht nach Klassen, sondern nach Lerngruppen differenziert wird. Unterrichtet eine Lehrkraft bspw. je eine achte Klasse in Englisch und Deutsch und schreibt mit beiden Klassen VERA, würde das dazu führen, dass die Grundgesamtheit der Lehrkräfte von der Anzahl der Lerngruppen abweicht und sich somit auf die Berechnung der Teilnahmequote auswirken. Ebenso ist zu beachten, dass bei der Anzahl der aufgerufenen Fragebögen grundsätzlich eine Mehrfachteilnahme einzelner Lehrkräfte bzw. ein mehrmaliges Aufrufen des Fragebogens durch dieselbe Person nicht ausgeschlossen werden kann. Jedoch können beide Sachverhalte aufgrund fehlender entsprechender Informationen nicht rekonstruiert werden. Somit sind die hier angegebenen Teilnahmequoten mit einiger Unsicherheit behaftet und daher mit Vorsicht zu interpretieren.

Tabelle 7 gibt einen Überblick über die in den einzelnen VERA-Durchgängen teilnehmenden Lerngruppen und die Häufigkeit, mit der die Befragung jeweils aufgerufen wurde, sowie die daraus resultierende Teilnahmequote. Im Jahr 2018 nahmen in den betrachteten Bundesländern insgesamt 15 052 Klassen aus 6 469 Schulen an VERA3 teil, während die Evaluationsbefragung 5 200 Mal aufgerufen wurde. Daraus ergibt sich eine Teilnahmequote von 35 %. Im darauffolgenden Jahr nahmen, u. a. durch das Ausscheiden bzw. die freiwillige Teilnahme des Landes Niedersachsen, nur 3 868 Schulen mit insgesamt 9 322 Klassen an der VERA3-Testung teil. Die Befragung wurde 3 644 Mal aufgerufen, was eine Rücklaufquote von 39 % bedeutet. Die Teilnahmebereitschaft an der Evaluation des VERA8-Durchgangs 2018 lag insgesamt deutlich unter der der beiden VERA3-Befragungen. An der Testung nahmen insgesamt 6 170 Lerngruppen aus 1 362 Schulen teil, der Fragebogen wurde nur 1 130 Mal aufgerufen. Die Teilnamequote unter den Sekundarschullehrkräften lag somit bei 18 %.

Tabelle 7: Teilnahmequoten der Datenerhebungen

	VERA3 2018	VERA3 2019	VERA8 2018
Anzahl Lerngruppen	15 052	9 322	6 170
Anzahl aufgerufener Fragebögen	5 200	3 644	1 130
Rücklaufquote	34.5 %	39.1 %	18.3 %

Die Teilnahmequoten sind zwar durchaus verbesserungswürdig, bewegen sich jedoch im erwartbaren Rahmen, da sie in etwa den in vergleichbaren Untersuchungen im Kontext von VERA erzielten Rücklaufquoten entsprechen (siehe bspw. Maier et al., 2012; Wacker & Kramer, 2012). Wacker und Kramer (2012) erzielen in einer Befragungsstudie zum ersten Messzeitpunkt eine Teilnahmequote von rund 27 %, während Maier et al. (2012) nur je nach Strichprobe zwischen 18 % und 26 % Rücklauf verzeichnen. Gerade im Kontext der Evaluationsbefragungen des zepf im Zuge der Durchführung der Vergleichsarbeiten erscheinen Rücklaufquoten zwischen 18 % und 39 % als durchaus zufriedenstellend. Groß Ophoff et al. (2019) sprechen bspw. von Teilnahmequoten von unter fünf Prozent aller teilnehmenden Schulen bei den Befragungen der Jahre 2012 bis 2015. Positiv ist zudem zu werten, dass die Rücklaufquoten zumindest in beiden VERA3-Durchgängen vergleichbar hoch sind.

Repräsentativität

Der Beschreibung der Rücklaufquote schließt sich die Frage nach der Repräsentativität der gewonnenen Stichproben an. Durch die freiwillige Teilnahme an den Befragungen ist die Thematik eines Nonresponse-Errors von Bedeutung. Aufgrund des dadurch entstehenden selbstselektiven Charakters der Datenerhebung bzw. der Stichproben können Erkenntnisse, die auf Basis solcher Erhebungen entstehen, nicht einfach generalisiert werden. Die Teilnahme kann durch persönliche Charakteristika beeinflusst sein, die sich auch auf die Ergebnisse der Untersuchung niederschlagen können. Im Hinblick auf diese Arbeit sind die Teilnehmenden möglicherweise positiver gegenüber Vergleichsarbeiten eingestellt und daher motivierter und eher bereit, den zusätzlichen Aufwand einer Befragung in Kauf zu nehmen. Als Gegenhypothese könnten es jedoch auch gerade diejenigen Lehrkräfte sein, die Vergleichsarbeiten besonders kritisch gegenüberstehen und in der Evaluationsbefragung ihrem Unmut Ausdruck verleihen wollen.

Dieses Problem eines möglichen Nonresponse-Errors kann zwar mit den vorliegenden Daten nicht beseitigt werden, jedoch können zumindest Erkenntnisse über mögliche Tendenzen im Antwortverhalten der untersuchten Stichprobe gewonnen werden. Im Zuge der VERA3-Testung werden in jedem Durchgang in den Ländern Bremen, Niedersachsen, Rheinland-Pfalz, Saarland, Schleswig-Holstein und Mecklenburg-Vorpommern neben den Evaluationsbefragungen sogenannte Zentralstichprobenbefragungen durchgeführt. Zweck dieser Befragung ist es, Kontextinformationen zur Gruppenbildung für die VERA-Rückmeldungen zu erhalten. Hierfür wird eine Zufallsstichprobe gezogen, wobei die Teilnahme an der Befragung für die ausgewählten Lehrkräfte verpflichtend ist. Durch die Zufallsauswahl in Kombination mit einer Teilnahmeverpflichtung kann bei dieser Befragung weitestgehend von einer Absenz eines Nonresponse-Errors und einer zumindest näherungsweise adäquaten Repräsentation der Grundgesamtheit ausgegangen werden (vgl. hierzu Koch, 2011).

Neben der Erhebung verschiedener Kontextinformationen, bspw. zur Zusammensetzung von Klassen, gab es in den Zentralstichprobenbefragungen der Jahre 2018 und 2019 einige weitere Fragen und dabei auch Überschneidungen mit Items der jeweiligen Evaluationsbefragungen. U. a. wurde auch im Fragebogen der Zentralstichprobe das Item AE5: „Ich befürworte die Durchführung der Vergleichsarbeiten.“ erhoben. Der Abgleich des Antwortverhaltens der Lehrkräfte in der jeweiligen Zentralstichprobe auf Basis dieses Items mit dem der Lehrkräfte der entsprechenden selbstselektiven Evaluationsstichprobe erlaubt zumindest Aussagen über eine mögliche positive oder negative Tendenz der Evaluationsstichprobe im Vergleich zur

Grundgesamtheit. Zur Überprüfung möglicher Unterschiede zwischen Evaluations- und Zentralstichprobe wurden jeweils für die Jahre 2018 und 2019 für das Item AE5 *t*-Tests gerechnet (siehe Tabelle 8).

Tabelle 8: Signifikanztests zu Gruppenunterschieden zwischen Evaluations- und Zentralstichprobe VERA3 2018 und 2019 für das Item AE5

	Gruppe	Gruppenstatistiken			Levene-Test		t-Test auf Mittelwertgleichheit			Effektstärke
		n	*M*	*SD*	*F*	*p*	*t*	*df*	*p*	*d*
VERA3 2018	Evaluation	2 736	2.45	0.94	7.43	.006	7.05 [a]	1304.8 [a]	<.001	0.29
	ZFB	794	2.18	0.93						
VERA3 2019	Evaluation	1 496	2.54	0.97	0.09	.769	5.53	1997	<.001	0.29
	ZFB	503	2.27	0.95						

Anmerkungen. *N* (zfb 2018) = 813, *N* (VERA3 2018, ohne NW) = 2 792, *N* (zfb 2019) = 517, *N* (VERA3 2019, ohne NW) = 1 516.
Item AE5: Ich befürworte die Durchführung der Vergleichsarbeiten (Skala: 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu*).
[a] Welch-t-Test für ungleiche Varianzen, da basierend auf dem Levene-Test die Hypothese der Varianzhomogenität verworfen wird.

Hierbei ist zu beachten, dass es im Land Nordrhein-Westfalen keine Zentralstichprobenbefragung gibt, da Kontextinformationen dort anderweitig zur Verfügung gestellt werden. Daher wurden auch in den Evaluationsstichproben die Lehrkräfte aus Nordrhein-Westfalen für diese Analyse ausgeschlossen, um eine bessere Vergleichbarkeit zu erzielen. Die Ergebnisse der t-Tests weisen aus, dass sich die Gruppen in den beiden betrachteten Jahren hinsichtlich des betrachteten Items jeweils signifikant unterscheiden ($p < .001$). Ein Blick auf die Mittelwertdifferenzen liefert Hinweise darauf, dass Lehrkräfte, die sich bereit erklären, an der Evaluationsbefragung teilzunehmen, den Vergleichsarbeiten grundsätzlich etwas positiver gegenüberstehen als die Lehrkräfte der repräsentativen Zentralstichprobe. Die Berechnung der jeweiligen Effektstärken nach Cohen (1988) belegt für beide Jahre mit $d = .29$ einen eher kleinen Effekt. Die Berechnung der Effektstärke erfolgte mit Hilfe des Online-Rechners von Hemmerich (2015) basierend auf t-Wert und Freiheitsgraden sowie den Stichprobengrößen. Diese Hinweise auf einen positiven Bias der an der Evaluation teilnehmenden Lehrkräfte im Vergleich zur

Grundgesamtheit sollten bei der Interpretation bzw. hinsichtlich einer möglichen Generalisierung der Ergebnisse berücksichtigt werden. Für VERA8 stehen keine vergleichbaren Daten zur Verfügung, vermutlich kann auch dort von einem ähnlich positiven Bias ausgegangen werden.

4.2.3. Strategie der Datensatzaufbereitung

Die Rohdaten der Onlinebefragungen enthielten zunächst noch viele Fälle, die zahlreiche bzw. auch ausschließlich fehlende Werte umfassten. Da bei Befragungsstudien immer von einer möglichen Achtlosigkeit einzelner Befragten beim Ausfüllen eines Fragebogens ausgegangen werden muss (z.B. Huang, Curran, Keeney, Poposki & DeShon, 2012; Meade & Craig, 2012; Niessen, Meijer & Tendeiro, 2016), wurden die Datensätze zunächst intensiv inspiziert, um nicht plausibles Antwortverhalten sowie Fälle mit einem Übermaß an fehlenden Werten und weiteren Auffälligkeiten zu identifizieren und zu eliminieren. Zu den fehlenden Werten wurden dabei sowohl Antwortverweigerungen gezählt als auch die Auswahl der Antwortoption „kann ich nicht beurteilen“. Durch den Einbezug dieser Antwortkategorie wurde der Ausschluss von fehlenden Werten zwar theoretisch etwas restriktiver, letztendlich wurden dadurch jedoch nur unwesentlich mehr Fälle aus der Analyse ausgeschlossen.

Die Aufbereitung der Daten erfolgte separat für jeden Teildatensatz, da sich u. a. die Gesamtanzahl der Variablen je nach Datensatz unterscheidet. Das gesamte Vorgehen soll hier kurz beschrieben werden: Zur Beurteilung der Datenqualität wurden verschiedene Personenstatistiken berechnet, die die Anzahl fehlender Werte sowie spezifische Bearbeitungsmuster erfassten. Basierend auf der ermittelten Gesamtzahl fehlender Werte pro Fall wurden gemäß den Standards der AAPOR (2016) jeweils diejenigen Fälle direkt aus der weiteren Analyse ausgeschlossen, bei denen mehr als 50 % aller Antworten als fehlend kodiert waren. Dieses Vorgehen entspricht auch der Strategie der Datenaufbereitung von Zlatkin-Troitschanskaia et al. (2016) oder Groß Ophoff (2013).

Nach dem Ausschluss von Fällen aufgrund einer zu großen Anzahl fehlender Angaben wurde die Bearbeitungszeit inspiziert und diejenigen Fälle aus dem Datensatz ausgeschlossen, bei denen eine im Verhältnis zur Länge des Fragebogens unrealistisch kurze Bearbeitungsdauer vorliegt. Eine zu kurze Bearbeitungszeit ist demnach ein Zeichen für eine zu achtlose Auseinandersetzung mit dem Fragebogen (Niessen et al., 2016). Zur Bestimmung eines Cutoff-Wertes schlagen Huang et al. (2012) eine minimale Bearbeitungszeit von 2 Sekunden je Item vor.

Dieser Wert wurde als Referenzwert herangezogen und zusätzlich wurde Zeit zum Lesen des Begrüßungstextes sowie der Bearbeitungshinweise einkalkuliert. Darüber hinaus wurden die einzelnen Befragungen von Testpersonen bearbeitet, um zu einer weiteren Einschätzung einer realistischen minimalen Bearbeitungsdauer zu gelangen. Basierend auf diesen Überlegungen wurde für jede Befragung eine plausible minimale Bearbeitungszeit geschätzt und alle Fälle, die diese Zeit unterschritten, pauschal aus dem Datensatz eliminiert. Die Schätzung der Bearbeitungszeit bezieht sich dabei auf den gesamten Fragebogen, da die reelle Bearbeitungsdauer nur für die gesamte Befragung ermittelt werden kann. Ein maximal zulässiger Zeitraum wurde nicht festgelegt, da sehr lange Bearbeitungszeiten auf ein fehlendes aktives Abschließen des Fragebogens zurückzuführen sein könnten, was jedoch keine Auswirkungen auf das Antwortverhalten hätte. Die minimal zulässige Bearbeitungsdauer je Befragung ist in Tabelle 9 dargestellt.

Die verbliebenen Fälle wurden weiter nach Auffälligkeiten des Antwortverhaltens, wie Hinweisen auf nachlässige Bearbeitung, untersucht. Hierzu wurden zur Beurteilung verschiedene Kennwerte ermittelt: Die Häufigkeit extremer und mittlerer Antwortkategorien, die maximale Folge der Auswahl derselben Antwortkategorie (*Maximum Longstring*) und die Personenstandardabweichung je Fall über alle betrachteten Items.

Die Betrachtung der Häufigkeit der Auswahl extremer und mittlerer Kategorien gibt Hinweise darauf, ob die Befragten zu extremem Antwortverhalten neigen oder eher eine Tendenz zur Mitte vorliegt. Sowohl bei der Tendenz zu extremen Antwortkategorien, als auch bei der Neigung, extreme Antwortoptionen eher zu vermeiden, handelt es sich um typische, relativ stabile Beantwortungsmuster, die häufig unabhängig vom Inhalt einer Befragung auftreten (Bachman & O'Malley, 1984; Emons, 2008). Der Wert der extremen Antwortkategorie ergibt sich aus der Summe der durch eine*n Proband*in gewählten extremen Antworten (Zijlstra, van der Ark & Sijtsma, 2011). Allein auf Basis eines besonders großen Anteils an Extremwerten kann jedoch nicht bestimmt werden, ob dies wirklich auf ein unsauberes Antwortverhalten zurückzuführen ist oder ein*e Teilnehmer*in einfach eine besonders starke Zustimmung oder Ablehnung zu den Items bzw. dem Untersuchungsgegenstand zum Ausdruck bringt. Für eine adäquate Bewertung sind daher weitere Informationen nötig (Zijlstra et al., 2011). In dieser Arbeit wurden für den Kennwert extremer Kategorien je Fall die Häufigkeit extremer Zustimmung bzw. Ablehnung über alle Items, also die Auswahl der Extremwerte 1 (*stimme überhaupt nicht zu*) und 4 (*stimme voll und ganz zu*) bzw. bei den Items zur zeitlichen Belastung 1 und 5 berechnet. Die

Häufigkeit mittlerer Kategorien ergibt sich analog dazu durch die Zählung der ausgewählten mittleren Antwortkategorien je Fall über alle Variablen.

Als weiterer Indikator für unsauberes und nachlässiges Antwortverhalten wurde der Maximum Longstring je Fall berechnet, ein Wert, der die maximale Häufigkeit einer von einer Person in einer Reihe gewählten Antwortkategorie anzeigt (Huang et al., 2012; Meade & Craig, 2012; Niessen et al., 2016). Hat ein*e Teilnehmer*in einer Befragung bspw. acht Mal hintereinander den Wert 3 ausgewählt und bei den restlichen Fragen immer abwechselnd 2 und 4, dann liegt der Wert des Maximum Longstrings für diesen Fall bei 8. Der Wert berechnet sich somit durch das Auszählen der in einer Reihe angekreuzten identischen Antwortkategorien. Der höchste Wert je Fall ergibt den Maximum Longstring. Es ist dabei nicht trivial, einen Cutoff-Wert festzulegen, ab welchem nicht mehr von einer sorgfältigen Bearbeitung gesprochen werden kann (Johnson, 2005). Zwar gibt es Vorschläge, alle Fälle mit einem Wert über 6 bzw. 14 aus einem Datensatz auszuschließen (Huang et al., 2012), dieser erscheint jedoch für diese Arbeit zu restriktiv, weil nicht pauschal ausgeschlossen werden kann, dass Befragte bewusst eine Entscheidung für die gleiche Antwortkategorie in Folge getroffen haben.

Da die einzelnen Kennzahlen separat betrachtet nur bedingt aussagekräftig sind, wurden die verschiedenen Indikatoren, wie bspw. Meade und Craig (2012) empfehlen, in Kombination zum Datenscreening herangezogen. Zusammen ergeben die beschriebenen Kennwerte erste Erkenntnisse über das Antwortverhalten und ermöglichen es, einzelne Fälle mit besonders auffälligem Antwortverhalten zu identifizieren und die Datensätze entsprechend zu bereinigen. Besonderes Augenmerk wurde dabei auf Fälle mit Auffälligkeiten bei der Beantwortung von Itemgruppen mit negativ gepolten Items gelegt. Würde bei einem Fall, der aufgrund eines besonders hohen Maximum Longstring Wertes bereits auffällig wurde, bei einer Itemgruppe mit einem oder mehreren negativen Items durchgängig ein Extremwert, z. B. 4, ausgewählt, wäre dies ein Hinweis auf ein unsauberes Antwortverhalten und der Fall würde aus dem Datensatz entfernt. Im folgenden Abschnitt ist die Datenaufbereitung je Datensatz beschrieben.

4.2.4. Datensatzaufbereitung

Die Aufbereitung der Datensätze erfolgte teils in Anlehnung an den Vorschlag der American Association for Public Opinion Research (AAPOR, 2016), bei der Beurteilung der Vollständigkeit einzelner Fälle eines Datensatzes nur die für die vorgesehenen Analysen relevanten Items

zu betrachten. Für diese Arbeit bedeutet dies Folgendes: Die Gesamtfragebögen umfassten neben den Items, die in die Auswertungen dieser Arbeit einflossen, weitere Fragenblöcke, die sich auf die Wahrnehmung von Lehrkräften zu VERA bezogen und für mögliche weitere Fragestellungen und Modellerweiterungen von Relevanz sein könnten. Diese wurden für die Datensatzaufbereitung beibehalten. Die Datensatzaufbereitung mit einem größeren Itemsample, als für die Beantwortung der eigentlichen Fragestellung notwendig, hat dabei mehrere Vorteile: In dem erweiterten Itemsample sind mehrere invers gepolte Items enthalten, die die Beurteilung eines möglichen inkonsistenten Antwortverhaltens erleichtern. Zusätzlich ergeben sich forschungs- und publikationslogische Vorteile, da sich weitere Anschlussuntersuchungen bzw. -publikationen wie bspw. Modellerweiterungen auf die (weitestgehend) gleichen Befragten beziehen, was die Kohärenz der Ergebnisse stärkt. Andere Items, die auch keinen erweiterten Bezug zur Fragestellung hatten, wie bspw. landesspezifische Fragen zu Informationsmaterialien wurden, ebenso wie Kontextinformationen zu Unterrichtsfach etc., für die Datensatzaufbereitung nicht berücksichtigt.

Der Umfang der in dieser Arbeit genutzten Befragungen, welche die Grundlage zur im Folgenden beschriebenen Aufbereitung der Datensätze bilden, ist in Tabelle 9 dargestellt. Der gesamte Fragebogen der VERA3-Evaluation 2018 bestand aus 86 Variablen, die für 5 200 Personen erhoben wurden. Davon wurden, wie beschrieben, die Items mit Kontextinformationen oder sonstigen für die empirische Auswertung nicht relevanten Informationen nicht mit in die Datensatzaufbereitung eingeschlossen. Nach Ausschluss dieser 25 Items verblieben 61 Variablen im Datensatz. Die VERA3-Evaluationsbefragung im Jahr 2019 umfasste 87 Variablen, von denen 18 ausgeschlossen wurden, sodass 69 Items für die Aufbereitung verblieben. Diese Befragung wurde von insgesamt 3 644 Personen aufgerufen. Der ursprüngliche Datensatz der VERA8-Evaluation 2019 beinhaltete 89 Variablen, die für 1 130 Personen erhoben wurden. 31 Variablen bezogen sich auf nicht relevante landesspezifische Inhalte oder reine Kontextinformationen und wurden daher ausgeschlossen, sodass 58 Variablen in die Analyse des Antwortverhaltens einflossen.

Im ersten Schritt der Datenaufbereitung wurden Muster fehlender Werte analysiert, für jeden Datensatz der kritische Wert gerade noch zulässiger fehlender Werte ermittelt (siehe Tabelle 9) und die Fälle, die diesen Grenzwert überschritten, aus den Daten entfernt (siehe Tabelle 10). Die Fälle, die aufgrund zu vieler fehlender Werte aus den Datensätzen ausgeschlossen wurden, werden in Tabelle 10 nach zwei Gruppen unterschieden. Zum einen hatten je nach Erhebung

zwischen 13 % und 16 % der Teilnehmenden den Fragebogen entweder nur aufgerufen und direkt wieder geschlossen oder hatten den Fragebogen (teilweise) durchgeklickt, ohne Fragen zu beantworten; diese Teilnehmenden werden Lurker genannt (Bosnjak & Tuten, 2001). Zum anderen hatten zwischen 2 % und 4 % der Befragten jeweils zumindest einige Fragen beantwortet, jedoch weniger als 50 %. Je nach Befragung hatten zwischen 14 % und 28 % der Teilnehmenden die jeweiligen Fragebögen vollständig ausgefüllt und alle betrachteten Items bearbeitet. Weitere 56 % bis 64 % der Proband*innen füllten die Befragungen zu mindestens 50 % aus. Nach der Analyse der fehlenden Werte verbleiben für die VERA3-Befragung 2018 4 273 Fälle (82 %) im Datensatz, für 2019 2 849 Fälle (78 %) und im Datensatz der VERA8-Befragung aus dem Jahr 2018 944 Fälle (84 %).

Tabelle 9: Umfang der Evaluationsbefragungen 2018 und 2019

	VERA3 2018	VERA3 2019	VERA8 2018
Anzahl Items gesamt	86	87	89
Anzahl Items für Datenbereinigung	61	69	58
Max. zulässige Anzahl Missings	30	34	29
Min. zulässige Bearbeitungszeit	5 min.[a] bzw. 5:30 min.	5:30 min.	5:30 min.

Anmerkungen. [a] Gilt nur für Rheinland-Pfalz; in der dortigen Befragung wurden 14 Items aufgrund ministerialer Vorgaben nicht erhoben. Diese 14 Items sind jedoch nicht in den 61 für die Datenbereinigung genutzten enthalten.

Die verbliebenen Fälle wurden hinsichtlich der Bearbeitungsdauer inspiziert und diejenigen mit einer zu kurzen Bearbeitungszeit, bezogen auf den gesamten Fragebogen, aus dem Datensatz ausgeschlossen, da eine sorgfältige Bearbeitung in einer zu kurzen Zeit in Frage gestellt werden muss. Bei den beiden VERA3-Erhebungen mussten jeweils weitere 3 % der Fälle aufgrund der Bearbeitungszeit aus den Daten entfernt werden, bei der VERA8-Evaluation 5 %.

Tabelle 10: Formen der Fragebogenbearbeitung

	VERA3 2018		VERA3 2019		VERA8 2018	
	n	%[a]	*n*	%[a]	*n*	%[a]
Aufrufe gesamt	5 200		3 644		1 130	
Non-Responder/Lurker	808	16 %	667	18 %	150	13 %
Einige Fragen beantwortet (<50 %)	119	2 %	128	4 %	36	3 %
hinreichend ausgefüllt (50-99 %)	3 255	63 %	2 337	64 %	632	56 %
vollständig ausgefüllt (100 %)	1 018	20 %	512	14 %	312	28 %
Gültige Fälle (nach Analyse der fehlenden Werte)	4 273	82 %	2 849	78 %	944	84 %
Ausschluss aufgrund von Bearbeitungsdauer [b]	129	3 %	97	3 %	53	5 %
Ausschluss aufgrund von Auffälligkeiten [b, c]	3	< 1 %	1	< 1 %	4	< 1 %
Gültige Fälle gesamt	4 141	80 %	2 751	75 %	887	78 %

Anmerkungen. [a] Aufgrund von Rundungsfehlern sind Abweichungen von 100 % möglich. [b] Fälle und Prozentwerte beziehen sich auf die nach den vorherigen Schritten verbliebenen Fälle (= Gültige Fälle (nach Analyse der fehlenden Werte)). [c] siehe auch Tabelle 11.

In einem weiteren Schritt folgte die Analyse der beschriebenen Bearbeitungsmuster mit den verbliebenen Fällen, die im Folgenden für jede Teilstichprobe separat beschrieben wird. Tabelle 11 enthält jeweils Minimum, Maximum, Median und Modus der jeweiligen Extremwerte, der Tendenz zur Mitte und des Maximum Longstring und wird im folgenden Abschnitt durch weitere Informationen ergänzt. Die Übersicht in Tabelle 11 verdeutlicht weitgehend vergleichbare Tendenzen in den Bearbeitungsmustern über die einzelnen Stichproben hinweg, auch unter Einbezug der unterschiedlichen Anzahl für die Analysen berücksichtigter Items.

Tabelle 11: Bearbeitungsmuster der Evaluationsbefragungen

Datensatz	Kennwert	Min.	Max.	Median	Modus
VERA3 2018 [a]	Extremwerte	0	58	14	10
	Tendenz zur Mitte	0	61	43	50
	Max Longstring	2	35	8	6
VERA3 2019 [b]	Extremwerte	0	64	18	10
	Tendenz zur Mitte	0	69	47	52
	Max Longstring	2	48	9	6
VERA8 2018 [c]	Extremwerte	0	52	15	8
	Tendenz zur Mitte	0	56	36	42
	Max Longstring	2	35	7	6

Anmerkungen. [a] 61 Items, n = 4 144; [b] 69 Items, n = 2 752; [c] 58 Items, n = 891.

In der Befragung der VERA3-Lehrkräfte des Jahres 2018 liegt die maximale Anzahl extremer Werte bei 58, während 31 Personen extreme Werte komplett vermieden. Der Modalwert liegt bei 10 extremen Werten (n = 183) und der Median bei 14. Insgesamt zeigt sich keine besonders starke Neigung zu extremem Antwortverhalten, wie auch die Abbildung 26 in Anhang A veranschaulicht. Bei der Inspektion der 50 Fälle mit der insgesamt stärksten Tendenz zu extremen Kategorien wird eine eher negative Tendenz im Antwortverhalten deutlich. Darüber hinaus werden keine besonderen Auffälligkeiten sichtbar, weshalb zunächst keine Fälle aufgrund extremen Antwortverhaltens ausgeschlossen wurden. Dagegen deutet das Antwortverhalten der Proband*innen auf eine relativ starke Tendenz zu mittleren Antwortkategorien hin (siehe Anhang A, Abbildung 27). Der Modus liegt bei 50 Werten der mittleren Antwortkategorie (n = 167), der Median bei 43. Ausschließlich mittlere Kategorien wurden von 7 Personen gewählt, nur 4 vermieden diese komplett. Hinsichtlich der Analyse des Maximum Longstring zeigt sich, dass keine Person ausschließlich einen Wert angekreuzt hat. Die maximale Anzahl gleicher fortlaufend gewählter Werte liegt bei 35. Der Modus dieses Kennwerts liegt bei 6 (n = 749), der Median bei 8. Ein Blick auf die in Abbildung 28 (Anhang A) dargestellte Verteilung verdeutlicht, dass nur einzelne Proband*innen über längere Strecken gleiche Werte angekreuzt haben und zunächst insgesamt keine starke Tendenz zu gleichen Antwortmustern zu erkennen ist. Abschließend wurden alle Fälle noch einmal auf Plausibilität inspiziert, um zu entscheiden, ob ggf. einzelne Fälle aus der weiteren Analyse ausgeschlossen werden sollten. Konkret wurden diejenigen Fälle herangezogen, die im Hinblick auf mehrere

Personenstatistiken Auffälligkeiten aufwiesen. Fälle mit vielen Extremwerten und einem hohen Maximum Longstring Wert wurden, vor allem mit Blick auf die 10 invers gepolten Variablen des Datensatzes, genauer betrachtet, da sich anhand dieser Variablen potenziell unsauberes Antwortverhalten identifizieren lässt. Daraufhin wurden 3 weitere Fälle aus dem Datensatz ausgeschlossen (siehe auch Tabelle 10).

Im Teildatensatz der VERA3-Evaluation des Jahres 2019 erweist sich eine Person mit 64 extremen Antworten als besonders auffällig im Antwortverhalten. Der Modalwert der extremen Werte liegt bei 10 (n = 109), der Median bei 18. 19 Personen vermieden die Auswahl extremer Antworten komplett. Insgesamt zeigt sich auch bei der Evaluation 2019 keine besonders starke Neigung zu extremem Antwortverhalten (siehe Anhang A, Abbildung 29). Die Inspektion der Fälle mit der stärksten Tendenz zu extremen Kategorien legt auch hier auf den ersten Blick ein eher konsistent negatives Antwortverhalten offen. Des Weiteren stechen keine besonderen Auffälligkeiten hervor, weshalb zunächst keine Fälle aufgrund extremen Antwortverhaltens ausgeschlossen wurden. Vergleichbar mit den Daten des Vorjahres ist auch im Jahr 2019 eine starke Tendenz zu mittleren Antwortkategorien erkennbar (siehe Anhang A, Abbildung 30). Der Modus liegt hierbei bei 52 (n = 99) gewählten mittleren Antworten und der Median bei 47. Fünf Personen haben ausschließlich mittlere Antwortkategorien gewählt, 4 Personen kreuzten keine mittleren Werte an. Bezüglich des Maximum Longstring gab es keine Person, die ausschließlich aufeinanderfolgend einen Wert angekreuzt hat. Die maximale Anzahl identischer in einer Reihe gewählten Werte liegt bei 48. Da es sich hierbei um denselben Fall handelt, der bereits 64 Extremwerte aufweist, wurde dieser aus dem Datensatz ausgeschlossen, auch da bei näherer Betrachtung deutlich wurde, dass die invers gepolten Items nicht adäquat erkannt wurden und somit nicht von einer sorgfältigen Bearbeitung ausgegangen werden kann. Alle weiteren Fälle mit höheren Werten dieses Indikators wurden zunächst beibehalten, da keine weiteren eindeutigen Hinweise auf unsaubere Bearbeitung des Fragebogens erkennbar waren. Das Minimum des Maximum Longstring liegt bei 2 aufeinanderfolgend gewählten identischen Werten, der Modus bei 6 (n = 498) und der Median bei 9. Auch hier macht ein Blick auf die Verteilung (siehe Anhang A, Abbildung 31) klar, dass nur einzelne Proband*innen über längere Strecken gleiche Werte angekreuzt haben und insgesamt keine starke Tendenz zu gleichen Antwortmustern zu erkennen ist. Die abschließende Inspektion anhand der Kombination der verschiedenen Kennzahlen ergab keine weiteren Auffälligkeiten, die eine Eliminierung weiterer Fälle nahegelegt hätte.

Analog zur VERA3-Evaluation wurden auch für die Stichprobe der VERA8-Lehrkräfte des Jahres 2018 Bearbeitungsmuster analysiert. In dieser Teilstichprobe liegt die maximale Anzahl gewählter extremer Werte bei 52 (n = 1). 5 Personen vermieden extreme Antwortkategorien vollständig. Der Modalwert liegt bei 8 extremen Werten (n = 49), der Median bei 15. Auch bei der VERA8-Evaluation 2018 ist keine besonders starke Neigung zu extremen Kategorien erkennbar (siehe Anhang A, Abbildung 32), ebenso wie bei VERA3 zeigt jedoch auch hier die Betrachtung der Fälle mit der größten Tendenz zu extremen Kategorien eine Neigung zu eher konsistent negativen als positiven Einschätzungen. Die Betrachtung der Häufigkeit der Auswahl mittlerer Kategorien vermittelt, ebenso wie bei VERA3, ein Bild von einer starken Tendenz zur Mitte (siehe Anhang A, Abbildung 33). Der Maximalwert liegt bei 56 (n = 1), 4 Personen vermieden mittlere Antwortkategorien komplett. Der Modalwert liegt hier bei 42 (n = 41), der Median bei 36 gewählten mittleren Antworten. Auch die Verteilung der Werte des Maximum Longstring ist vergleichbar mit derjenigen der VERA3-Daten (siehe Anhang A, Abbildung 34). Die maximale Anzahl identischer aufeinanderfolgend gewählter Werte liegt bei 35 (n = 1), das Minimum, ebenso wie bei den anderen Teildatensätzen, bei 2. Der Modus liegt bei 6 in einer Reihe gewählter gleicher Antwortkategorien (n = 184), der Median bei 7. Wie bei den anderen Datensätzen wurden basierend auf einer gesamtheitlichen Betrachtung der Kennwerte einzelne Fälle näher betrachtet und vier Fälle daraufhin aus dem Datensatz entfernt, da deren Antwortmuster Rückschlüsse auf eine unsaubere Bearbeitung der Befragung zuließen.

Insgesamt ist das Antwortverhalten in allen drei Teilstichproben grundsätzlich vergleichbar und weist ähnliche Tendenzen auf. Hinsichtlich der Bereinigung der Datensätze sei angemerkt, dass hierbei nicht allzu restriktiv vorgegangen wurde, um, analog zur Kritik am Vorgehen eines listenweisen Fallausschlusses bei fehlenden Werten, mögliche Verfälschungen der Ergebnisse sowie den Verlust statistischer Power durch den Ausschluss zu vieler Fälle aus den Daten (Engel & Schmidt, 2014; Graham, Cumsille & Elek-Fisk, 2003) zu vermeiden. Es wurden nur wenige Fälle aufgrund von Auffälligkeiten in den verschiedenen Kennzahlen eliminiert, da eine Unterscheidung zwischen tatsächlich nachlässigem oder absichtlich verfälschendem Antwortverhalten und ernsthafter Beantwortung des Fragebogens anhand der Indikatoren nur schwer zu treffen ist. Bspw. kann es gerade bei einer starken Tendenz zu mittleren Antwortkategorien durchaus sein, dass bei einer Fragengruppe mit vier Items, bei denen eines invers gepolt ist, bei allen Fragen bewusst die Antwortoption 3 gewählt wurde, ohne dass eine Nachlässigkeit des*der Teilnehmer*in vorlag. Die Indikatoren dienen somit in erster Linie einem ersten Kennenlernen der Daten sowie als mögliche spätere Interpretationshilfen der Ergebnisse.

4.2.5. Finale Stichproben

Stichproben der Modelltestung und Validierung

Datengrundlage zur Überprüfung des Forschungsmodells (siehe Kapitel 5.1) und der Modellvalidierung bzw. der Validierung eines angepassten Modells (siehe Kapitel 5.3) bilden die aufbereiteten VERA3-Evaluationsbefragungen der Jahre 2018 und 2019. Die final aufbereitete VERA3-Stichprobe des Jahres 2018, auf deren Basis das in Kapitel 3 aufgestellte Forschungsmodell und die damit einhergehenden Hypothesen getestet wurden, umfasst insgesamt 4 144 Fälle. Die Datengrundlage der Validierungsstudie auf Basis der VERA3-Evaluationsbefragung im Jahr 2019 umfasst nach der Datenaufbereitung 2 751 Fälle. In beiden Teilstichproben nahmen Lehrkräfte aus Bremen, Niedersachsen, Rheinland-Pfalz, Saarland, Schleswig-Holstein, Nordrhein-Westfalen und Mecklenburg-Vorpommern teil, im Jahr 2019 zusätzlich noch einige Lehrkräfte der Deutschsprachigen Gemeinschaft in Belgien. Die deutlich geringere Zahl der Teilnehmenden im Jahr 2019 im Vergleich zum Vorjahr lässt sich durch die Entscheidung des Landes Niedersachsen erklären, nicht mehr an den Vergleichsarbeiten teilzunehmen, bzw. die Verpflichtung zur Teilnahme im Jahr 2019 zunächst aufzuheben.

Tabelle 12: Verteilung der Testfächer VERA3 2018 und 2019

		Nur Mathematik	Nur Deutsch	Beide Fächer
VERA3 2018	Abs. H.	685	1 450	1 988
	Rel. H.	16.6	35.2	48.2
VERA3 2019	Abs. H.	497	584	1 663
	Rel. H.	18.1	21.3	60.6

Anmerkungen. *N* (VERA3 2018) = 4 141; *N* (VERA3 2019) = 2 751; Die Frage nach den Testfächern hatte ein Mehrfachantwortformat; Abs. H.: Absolute Häufigkeit, Rel. H.: Relative Häufigkeit jeweils in Prozent.

Aus datenschutzrechtlichen Gründen konnten in den Befragungen kaum Hintergrundinformationen zu den teilnehmenden Lehrkräften erhoben und daher nur wenige Aussagen zu den Eigenschaften der Stichprobe getroffen werden. Es wurden lediglich die Fächer, in denen VERA geschrieben wurde, erfasst (siehe Tabelle 12). In beiden Stichproben liegt der Anteil an Lehrkräften, die nur Mathematik getestet hatten, mit 16.6 % bzw. 18.1 % am niedrigsten. Der Anteil derjenigen Lehrkräfte, die ausschließlich Deutsch testeten, liegt im Durchgang 2019 nur geringfügig höher bei 21.3 %, im Jahr 2018 bei immerhin 35.2 %. In beiden Gruppen nahm mit 48.2 % bzw. 60.6 % die Mehrheit der Lehrkräfte in beiden Fächern an VERA3 teil.

Stichproben der latenten Zweigruppenanalyse

Grundlage der latenten Gruppenanalyse (siehe Kapitel 5.4) bilden die Daten der VERA8-Evaluationsbefragung 2018 und eine Teilstichprobe der VERA3-Befragung 2018. Die Stichprobe der Evaluationsbefragung im Zuge der VERA8-Testung im Jahr 2018 umfasste nach der Datenbereinigung noch insgesamt 887 Fälle. Hierunter fallen Lehrkräfte aus den Ländern Bremen, Niedersachsen, Rheinland-Pfalz und dem Saarland.

Während die Vergleichsarbeiten in der Gruppe der Grundschullehrkräfte ausschließlich papierbasiert durchgeführt wurden, umfasst die hier beschriebene Stichprobe der VERA8-Lehrkräfte einerseits Lehrkräfte, die mit ihrer Lerngruppe die Vergleichsarbeiten ebenfalls papierbasiert (PP) durchgeführt haben, und andererseits auch Lehrkräfte, die an einer computerbasierten Testung (CBT) teilnehmen konnten. Da die Vermutung nahe liegt, dass sich die beiden Gruppen von Lehrkräften in ihrer Bewertung der Vergleichsarbeiten zumindest hinsichtlich einzelner Aspekte unterscheiden, wurden verschiedene t-Tests gerechnet, um mögliche Gruppenunterschiede hinsichtlich des Durchführungsmodus aufzudecken und zu entscheiden, ob der Teil der Lehrkräfte, der computerbasiert getestet hatte, aus der Gruppenanalyse (VERA3 vs. VERA8) ausgeschlossen werden sollte. Die Ergebnisse der t-Tests, einschließlich Levene-Tests nach Levene (1960) auf Varianzgleichheit zur Prüfung der Voraussetzung eines herkömmlichen t-Tests, sind ausführlich in Tabelle 13 dargestellt. Zunächst wurden die manifesten Konstruktmittelwerte auf signifikante Mittelwertunterschiede zwischen der CBT- und der PP-Gruppe untersucht. Auf Konstruktebene zeigen sich keine signifikanten Unterschiede im Hinblick auf den Testdurchführungsmodus. Auch die jeweiligen Effektstärken nach Cohen weisen mit maximal 0.17 nicht auf substanzielle Gruppeneffekte hin.

Zusätzlich wurden jedoch die Items des Konstrukts zeitliche Belastung (ZB1, ZB2, ZB3) auf signifikante Mittelwertunterschiede getestet, da eine computerbasierte Testdurchführung zu einem reduzierten Arbeitsaufwand im Vergleich zur herkömmlichen papierbasierten Testung führen sollte und daher zumindest bei einzelnen Items durchaus Unterschiede zu erwarten wären. Wie Tabelle 13 zu entnehmen ist, zeigen sich für Item ZB1 und ZB3 signifikante Mittelwertunterschiede. Auch die Effektstärke d weist bei Item ZB1 auf einen kleinen bis mittleren (d = 0.38), bei Item ZB3 sogar auf einen starken (d = -0.86) Gruppeneffekt hin. Um mögliche Verzerrungen bei der Modellschätzung zu vermeiden, wurden daher die Fälle, die computerbasiert getestet hatten, aus der Analyse ausgeschlossen. Die Größe der Stichprobe der VERA8-Lehrkräfte für die Gruppenanalyse beträgt daher N = 782.

Tabelle 13: Signifikanztests zu Gruppenunterschieden zwischen CBT und PP Testung (VERA8)

Konstrukt/ Item	Gruppe	Gruppenstatistiken			Levene-Test		t-Test auf Mittelwert-gleichheit			Effekt-stärke
		n	*M*	*SD*	*F*	*p*	*t*	*df*	*p*	*d*
NI [a]	CBT	96	2.32	0.68	1.61	.205	1.56	824	.119	0.17
	PP	730	2.20	0.71						
AE [a]	CBT	100	2.34	0.81	0.27	.601	0.01	843	.992	0.00
	PP	745	2.34	0.84						
WN [a]	CBT	82	2.30	0.75	0.90	.344	-0.11	772	.910	-0.01
	PP	692	2.31	0.78						
ZB [b]	CBT	98	3.00	0.72	0.44	.507	-1.48	851	.141	-0.16
	PP	755	3.11	0.68						
ZB1 [b]	CBT	100	2.84	1.21	2.49	.115	3.58	857	< .001	0.38
	PP	759	2.43	1.04						
ZB2 [b]	CBT	103	3.02	0.84	0.11	.740	0.04	871	.966	0.00
	PP	770	3.02	0.85						
ZB3 [b]	CBT	103	3.12	0.86	4.45	.035	-8.16	131.45	< .001	-0.86
	PP	778	3.88	0.88						

Anmerkungen. NI: Nutzungsintention, AE: Einstellung, ZB: Zeitliche Belastung, WN: Nützlichkeit.
CBT: Computerbasierte Testung, PP: Paper-Pencil-Testung;
N (VERA8 2018, CBT) = 105, *N* VERA8 2018, PP) = 782;
[a] Wertebereich der Variablen jeweils 1 bis 4; Min = 1. Max = 4. [b] Wertebereich der Variablen jeweils 1 bis 5; Min = 1. Max = 5 (inverse Polung). [c] Welch-t-Test für ungleiche Varianzen, da basierend auf dem Levene-Test die Hypothese der Varianzhomogenität verworfen wird.

Auch in der VERA8-Evaluationsbefragung wurden die jeweiligen Testfächer der Lehrkräfte erfragt. In VERA8 wurden im Jahr 2018 Mathematik, Deutsch, Englisch und in einigen Ländern auch Französisch getestet. Tabelle 14 gibt einen Überblick über die Testfächer der an der Befragung teilnehmenden Lehrkräfte, nach Ausschluss der CBT-Fälle. Die Mehrheit der Lehrkräfte (77.1 %) führte VERA ausschließlich im Fach Mathematik durch, für das die Durchführung 2018 in allen betrachteten Ländern verpflichtend war. Mit 10.4 % bzw. 7.9 % nahmen deutlich weniger Lehrkräfte der betrachteten Stichprobe ausschließlich an der Deutsch- bzw.

Englischtestung teil. 3.5 % der Befragten nahmen in mehreren Fächern an dem VERA8-Durchgang teil, 1.2 % machten keine Angabe zu einem Testfach.

Tabelle 14: Verteilung der Testfächer VERA8 2018

		Nur Mathematik	Nur Deutsch	Nur Englisch	Mehrere Fächer	Keine Angabe
VERA8 2018	Abs. H.	603	81	62	27	9
	Rel. H.	77.1	10.4	7.9	3.5	1.2

Anmerkungen. *N*(VERA8 2018) = 782; Die Frage nach den Testfächern hatte ein Mehrfachantwortformat; Abs. H.: Absolute Häufigkeit, Rel. H.: Relative Häufigkeit in Prozent.

Aus Gründen der Vergleichbarkeit wurden auch aus der VERA3-Stichprobe Fälle ausgeschlossen. Da in der VERA8-Stichprobe nur Daten der Bundesländer Bremen, Niedersachsen, Rheinland-Pfalz und Saarland vorliegen, wurden entsprechend aus dem VERA3-Datensatz alle Fälle von Lehrkräften aus Schleswig-Holstein, Nordrhein-Westfalen und Mecklenburg-Vorpommern für die Mehrgruppenanalysen ausgeschlossen. Die Stichprobengröße der VERA3-Lehrkräfte liegt somit bei $N = 1\,918$, die gesamte analysierte Stichprobe liegt bei $N = 2\,700$ (VERA3- und VERA8-Stichprobe). Die unterschiedliche Größe der Gruppen sollte, wie in Kapitel 4.3.4 beschrieben, keinen Einfluss auf die Schätzung des Mehrgruppenmodells nehmen (Koh & Zumbo, 2008; Schwab & Helm, 2015).

4.3. Analysemethoden

4.3.1. Auswahl der Analysemethode: Strukturgleichungsmodellierung zur Untersuchung der theoretischen Wirkungszusammenhänge

Zur Untersuchung der theoretischen Wirkungszusammenhänge des aufgestellten Forschungsmodells wurde das Verfahren der *Strukturgleichungsmodellierung (SEM - Structural Equation Modelling)* gewählt. Der Begriff SEM steht nicht für eine einzelne Methode, sondern für eine ganze Gruppe verschiedener Techniken multivariater, statistischer Datenanalysen, wie Faktorenanalysen, Pfad- oder Regressionsanalysen (Kline, 2011; Reinecke, 2014). Entscheidend für diese Arbeit ist die Eignung von SEM-Analysen für die statistische Untersuchung theoretisch begründeter, auch komplexer Wirkungszusammenhänge (Ullman & Bentler, 2003; Urban &

Mayerl, 2014), um eine empirische Überprüfung des in Kapitel 3 entwickelten theoretischen Modells einschließlich der aufgestellten Hypothesenstruktur zu ermöglichen.

In einem Strukturgleichungsmodell werden zunächst zwei verschiedene Variablentypen unterschieden: manifeste und latente Variablen. Eine latente Variable repräsentiert ein theoretisches Konstrukt, bzw. einen Faktor, welches nicht direkt beobachtbar und somit nicht direkt messbar ist, wie die in dieser Arbeit untersuchten Einstellungen und individuellen Wahrnehmungskonstrukte. Diese latenten Konstrukte werden wiederum von, meist mehreren, manifesten Variablen repräsentiert. Manifeste Variablen, i. d. R. Fragebogenitems, sind direkt messbar und dienen als Indikatoren der latenten Variablen (Kline, 2011; MacCallum & Austin, 2000). Manifeste Variablen können jegliche Ausprägungen (kategorial, ordinal oder stetig) haben, latente Variablen hingegen sind i. d. R. stetig (Kline, 2011). Die Kausalstruktur zwischen diesen manifesten Indikatorvariablen und den latenten Faktoren wird durch *Messmodelle*, *auch äußere Modelle* genannt, abgebildet. Diese postulierten Beziehungen zwischen latenter Variable und den bestimmenden Indikatoren werden typischerweise mithilfe von *konfirmatorischen Faktorenanalysen (CFA – Confirmatory Factor Analysis)* empirisch überprüft (Monecke & Leisch, 2012; Steinmetz, 2015). Messmodelle können entweder *reflektiven* oder *formativen* Charakter haben. Die Logik dieser beiden Spezifikationsarten wird in Kapitel 4.3.2 in den Abschnitten zur Spezifikation des Messmodels (siehe S. 134ff.) im Detail erläutert.

Die Beziehung der latenten Variablen untereinander wird innerhalb eines *Strukturmodells*, auch als *inneres Modell* bezeichnet, spezifiziert und mittels (latenter) Pfadanalyse geschätzt (Berning, 2019). Bei der Spezifikation multipler Abhängigkeiten wird dabei zwischen *exogenen* und *endogenen* Variablen unterschieden. Exogene Variablen werden im Modell durch keine anderen latenten Variablen erklärt, alle anderen Variablen, die durch weitere latente Variablen erklärt werden, sind endogen, auch wenn diese wiederum weitere Variablen beeinflussen (Berning, 2019). Der Aufbau von Struktur- und Messmodellen ist im Kapitel 4.3.2 ausführlicher dargestellt.

Neben der Konzeptualisierung latenter Konstrukte besteht eine weitere Besonderheit von Strukturgleichungsmodellen in der Berücksichtigung eines dritten Variablentyps, den *Residuen* oder Fehlertermen. Diese können entweder mit den beobachteten (Indikator-)Variablen oder den im Strukturmodell endogenen (abhängigen) latenten Faktoren assoziiert sein. Hinsichtlich der Indikatorvariablen repräsentiert der Fehlerterm den Varianzanteil, der nicht durch den zugrundeliegenden Faktor erklärt werden kann. Ein Teil dieser unerklärten Varianz entsteht dabei durch

zufällige Messfehler. Die Berücksichtigung dieser Messfehler in der Analyse stellt eine Stärke von Strukturgleichungsmodellen dar (Kline, 2011). Gerade in den Sozialwissenschaften kommen die Vorteile dieses zentralen Merkmals von SEMs zum Tragen, da gängige Messungen wie bspw. Befragungen i. d. R. nicht fehlerfrei durchführbar sind (Reinecke, 2014). In Survey-Studien zur Einstellungsmessung, wie sie auch in dieser Arbeit durchgeführt wurde, werden häufig eher spontane Meinungsäußerungen, basierend auf subjektiven Wissenselementen oder Bewertungen mit nur geringer kognitiver Zentralität und Stabilität, als konsistente Einstellungen erfasst. Derartige durch aktuelle Befindlichkeiten beeinflusste Momentaufnahmen können zu fehlerhaften Einstellungsmessungen und entsprechend fehlerhaften Effektschätzungen führen. Diesen messtheoretischen Einschränkungen begegnen Strukturgleichungsmodelle durch die Berücksichtigung von Messfehlern in einem fehlerkorrigierten Schätzverfahren, das fehlerbereinigte Schätzwerte liefert (Urban & Mayerl, 2014). Diese Berücksichtigung von Messfehlern auf Messmodellebene ermöglicht auf der Strukturebene eine messfehlerbereinigte Analyse der Beziehungen zwischen den latenten Konstrukten mit unverzerrten Schätzungen von Korrelations- und Regressionskoeffizienten (Ullman & Bentler, 2003; Weiber & Mühlhaus, 2014).

4.3.2. Aufbau von Strukturgleichungsmodellen

Grafische Darstellung eines Strukturgleichungsmodells

Da Strukturgleichungsmodelle mit zunehmender Anzahl manifester Variablen und zugehöriger Indikatorvariablen schnell sehr komplex werden können, empfiehlt sich i. d. R. eine grafische Darstellung der spezifizierten Zusammenhänge. Abbildung 9 veranschaulicht daher exemplarisch ein typisches Strukturgleichungsmodell gemäß der üblichen Darstellungskonventionen, Tabelle 15 fasst die zugehörigen Notationen zusammen. Das abgebildete Strukturgleichungsmodell beschreibt ein Messmodell für ein exogenes Konstrukt (ξ_1), die Messmodelle zweier endogener Konstrukte (η_1 und η_2) sowie die Beziehung dieser latenten Konstrukte untereinander ($\gamma_{11}, \gamma_{21}, \beta_{21}$). In der grafischen Darstellung repräsentieren Ovale die latenten Konstrukte, die Rechtecke die jeweiligen dazugehörigen Indikatorvariablen. Pfeile mit einer Pfeilspitze stellen gerichtete kausale Beziehungen dar, ungerichtete Pfeile mit zwei Spitzen ungerichtete korrelative Beziehungen (Ho, Stark & Chernyshenko, 2012).

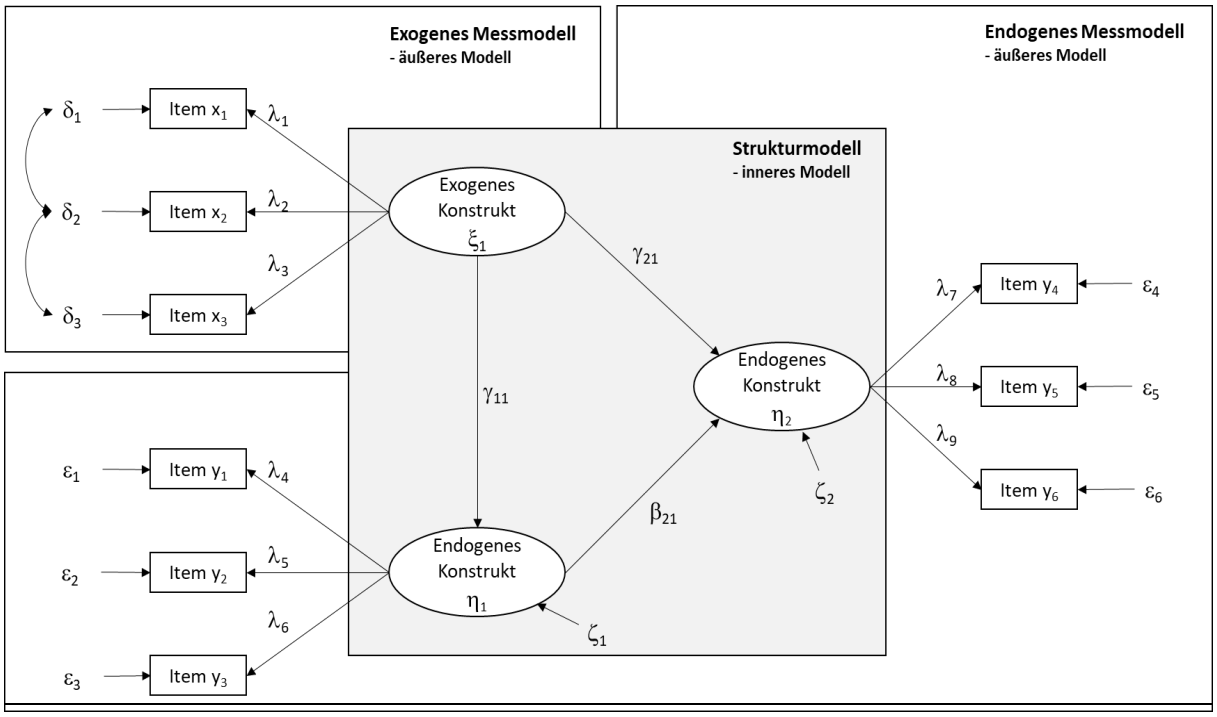


Abbildung 9: Aufbau eines typischen Strukturgleichungsmodells mit drei latenten Faktoren (eigene Darstellung in Anlehnung an Bollen, 2014; A. Fuchs, 2011)

Tabelle 15: Notationen von Strukturgleichungsmodellen

Symbol	Matrix	Definition
ξ (ksi)		Latente exogene Variable
η (eta)		Latente endogene Variable
λ (lambda)	Λ (Lambda)	Faktorladung
ζ (zeta)		Messfehler/Residualvariable für latente exogene Variablen
δ (delta)	Θ_δ (Theta delta)	Residualvariable für eine Indikatorvariable x bzw. Kovarianzmatrix von δ
ε (epsilon)	Θ_ε (Theta epsilon)	Residualvariable für eine Indikatorvariable y bzw. Kovarianzmatrix von ε
x		Indikatorvariable für eine latente exogene Variable
y		Indikatorvariable für eine latente endogene Variable
γ (gamma)		Regressionsgewicht assoziiert mit einer latenten exogenen Variablen
β (beta)		Regressionsgewicht assoziiert mit einer latenten endogenen Variablen
	Ψ (Psi)	Kovarianzmatrix der Fehlerterme der endogenen Variablen
	Φ (Phi)	Kovarianzmatrix der latenten exogenen Variablen

Das Strukturmodell

Das hier als inneres Modell gekennzeichnete Strukturmodell spezifiziert die Beziehungen zwischen den latenten exogenen (ξ_n) und endogenen (η_n) Konstrukten. (Standardisierte) Pfadkoeffizienten (γ_n bzw. β_n) quantifizieren, analog zu (standardisierten) Regressionskoeffizienten, die gerichtete Beziehung zwischen zwei latenten Variablen. Sie beschreiben das Ausmaß der Veränderung einer latenten abhängigen Variablen bei einer Veränderung der unabhängigen latenten Variablen um eine (Standard-)Einheit unter Konstanthaltung der anderen latenten unabhängigen Variablen (Bollen, 2014; Urban & Mayerl, 2014). Ein allgemeines Strukturmodell wird durch die folgende Gleichung beschrieben:

$$\eta = B\eta + \Gamma\xi + \zeta$$

In dieser Gleichung beschreibt Eta (η) den Vektor der endogenen latenten Variablen und Xi (ξ) den Vektor der exogenen latenten Variablen. Der Vektor Zeta (ζ) repräsentiert die Fehlerterme der exogenen Variablen. Die Matrix Beta (B) beinhaltet die Koeffizienten der latenten endogenen Variablen, also die Beziehungen der endogenen Konstrukte untereinander. Die Matrix Gamma (Γ) repräsentiert die Koeffizienten der Beziehungen zwischen exogenen und endogenen Variablen. Des Weiteren werden mit Phi (Φ) die Kovarianzmatrix der latenten exogenen Variablen und mit Psi (Ψ) die Kovarianzmatrix der Fehlerterme der endogenen Variablen spezifiziert (Berning, 2019; Bollen, 2014; Ho et al., 2012).

Indirekte Effekte

Strukturgleichungsmodelle erlauben nicht nur die Untersuchung direkter linearer Zusammenhänge, sondern auch die Analyse vermittelter Zusammenhänge und die Prüfung damit einhergehender verbundener Hypothesen. Wird der Zusammenhang zwischen zwei Konstrukten über mindestens ein drittes Konstrukt, eine Mediatorvariable, vermittelt, spricht man von einem indirekten Effekt (Berning, 2019). Ein indirekter Effekt berechnet sich als das Produkt der einzelnen Pfadkoeffizienten. Der Gesamteffekt ergibt sich aus der Summe der indirekten und direkten Effekte (Bollen, 2014). So bemisst sich bspw. der Gesamteffekt von ξ_1 auf η_2 folgendermaßen:

$$\begin{aligned}\text{Gesamteffekt} &= \text{direkter Effekt} + \text{indirekte Effekte} \\ &= \gamma_{21} + \gamma_{11}\beta_{21}\end{aligned}$$

In komplexeren Modellen lassen sich nicht nur einfache Mediationen wie der Pfad $\gamma_{11}\beta_{21}$ abbilden, sondern auch komplexere Wirkungszusammenhänge wie parallele oder serielle Mediationen. Diese Art der Modellierung, bei der eine Mediation durch eine (oder mehrere) weitere Variable vermittelt wird, ermöglicht es mit einem SEM Pfadabhängigkeiten zu untersuchen (Berning, 2019).

Spezifikation des Messmodells: reflektive vs. formative Modellierung

Das äußere Modell eines Strukturgleichungsmodells, das Messmodell, beschreibt die Beziehung zwischen einem latenten Konstrukt (ξ_n bzw. η_n) und dessen Indikatorvariablen (x_i bzw. y_i). Die in Abbildung 9 dargestellten Messmodelle folgen einer reflektiven Modellspezifikation. Einem reflektiven Messmodell liegt die Logik zugrunde, dass die Ausprägungen der beobachtbaren Indikatoren durch den zugrundeliegenden latenten Faktor erklärt werden. Eine Veränderung des Konstrukts spiegelt sich gemäß dieser Kausalitätsannahme in der Veränderung aller Indikatoren wider, wird also reflektiert. Ein latentes Konstrukt wird bei einem reflektiven Messmodell als eine Funktion seiner latenten Indikatoren modelliert (Christophersen & Grape, 2009). Abbildung 10 visualisiert ein solches reflektives Messmodell.

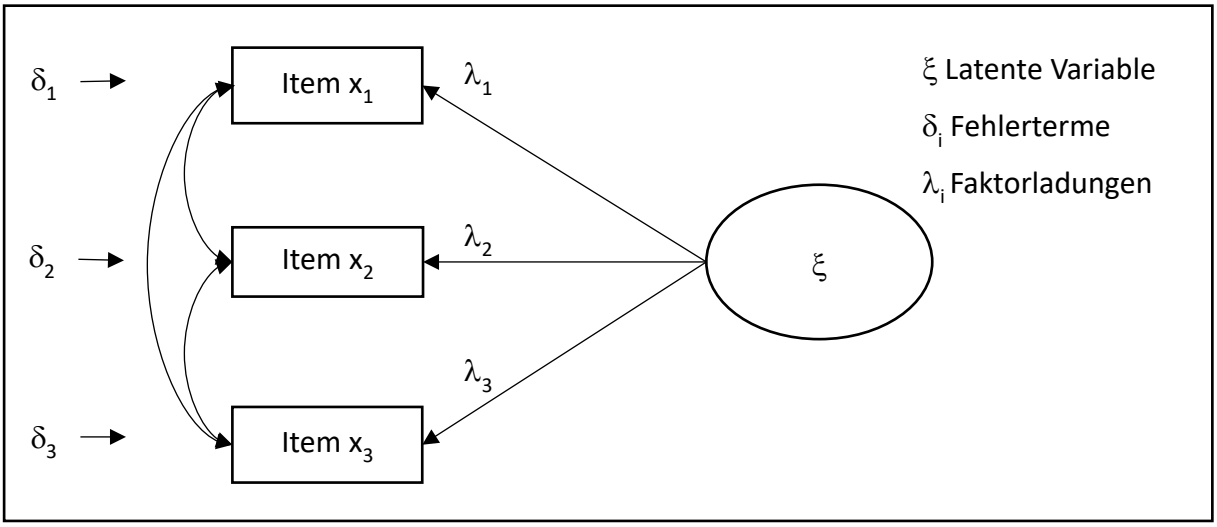


Abbildung 10: Reflektives Messmodell (eigene Darstellung in Anlehnung an Edwards, 2011, S. 372)

Gemäß einer reflektiven Modellierung wird die Messung eines jeden Indikators (x_i bzw. y_i) durch das zugrundeliegende Konstrukt (ξ_n bzw. η_n) und einen zufälligen Messfehler (δ_i bzw. ε_i) beeinflusst. Der zugrundeliegende Faktor ist für alle seine Indikatoren identisch, der Fehlerterm dagegen ist individuell bzw. itemspezifisch. Die Kovarianz zwischen den einzelnen

Indikatoren wird dabei deren gemeinsame Ursache, dem latenten Faktor, zugeschrieben (Edwards & Bagozzi, 2000). Die Indikatorvariablen eines Konstrukts sind bei einer reflektiven Modellierung i. d. R. hoch miteinander korreliert. Wäre eine fehlerfreie Messung möglich ($\delta_i = 0$), läge eine perfekte Korrelation zwischen den Indikatoren vor (Eberl, 2004).

In einer typischen CFA zur Schätzung eines reflektiven Messmodells ergibt sich der Wert eines Items daher aus der linearen Funktion einer latenten Variablen und einem stochastischen Fehlerterm (Zumbo, 2005). Mathematisch lassen sich reflektive Messmodelle somit folgendermaßen darstellen (Bollen, 2014):

$$x = \Lambda_x \xi + \delta$$
$$y = \Lambda_y \eta + \varepsilon$$

Xi (ξ) und Eta (η) bezeichnen die Vektoren der latenten exogenen bzw. endogenen Variablen. Die lambda-Matrizen Λ_x und Λ_y beinhalten die Koeffizienten lambda (λ_i), durch die der Einfluss des latenten Konstrukts auf seine Indikatorvariablen beschrieben wird. Diese sind als Regressionskoeffizienten der Effekte der latenten Variable auf die manifesten Indikatoren zu verstehen. Sie beschreiben das Ausmaß der Veränderung der Indikatorvariablen (x_i bzw. y_i) bei einer Veränderung der latenten Variablen um eine Einheit (Bollen, 2014). Delta (δ) und epsilon (ε) repräsentieren die Vektoren der exogenen und endogenen *Residualvariablen*, also die Messfehler der jeweiligen Indikatorvariablen x_i bzw. y_i (Bollen, 2014; A. Fuchs, 2011). Die Matrizen Theta delta Θ und Theta epsilon Θ beschreiben die Kovarianzmatrizen der Fehlerterme (Bollen, 2014).

Ein formatives Messmodell folgt gegenüber der reflektiven Modellierung einer umgekehrten Kausalitätsannahme. Ein latentes Konstrukt wird demnach ursächlich durch die Ausprägungen seiner manifesten Indikatoren bestimmt. Eine Veränderung eines oder mehrerer Indikatorvariablen führt somit zu einer Veränderung des latenten Faktors. Mathematisch ergibt sich ein formativ modelliertes latentes Konstrukt aus der gewichteten Zusammensetzung seiner Indikatoren (Christophersen & Grape, 2009):

$$\eta = \sum \gamma_i x_i + \zeta$$

Abbildung 11 visualisiert den entsprechenden Aufbau eines formativen Messmodells. Eta (η) repräsentiert dabei das latente Konstrukt, γ_i den Effekt der formativen Indikatorvariablen x_i auf

η. Die Darstellung verdeutlicht zudem, dass die Indikatorvariablen eines formativen Messmodells, im Gegensatz zu reflektiven Indikatoren, keine individuellen Fehlerterme besitzen. Der Fehlerterm Zeta (ζ) wird stattdessen bei der latenten Variablen spezifiziert und repräsentiert Aspekte der latenten Variablen, die nicht durch die Indikatorvariablen erklärt werden (Edwards, 2011).

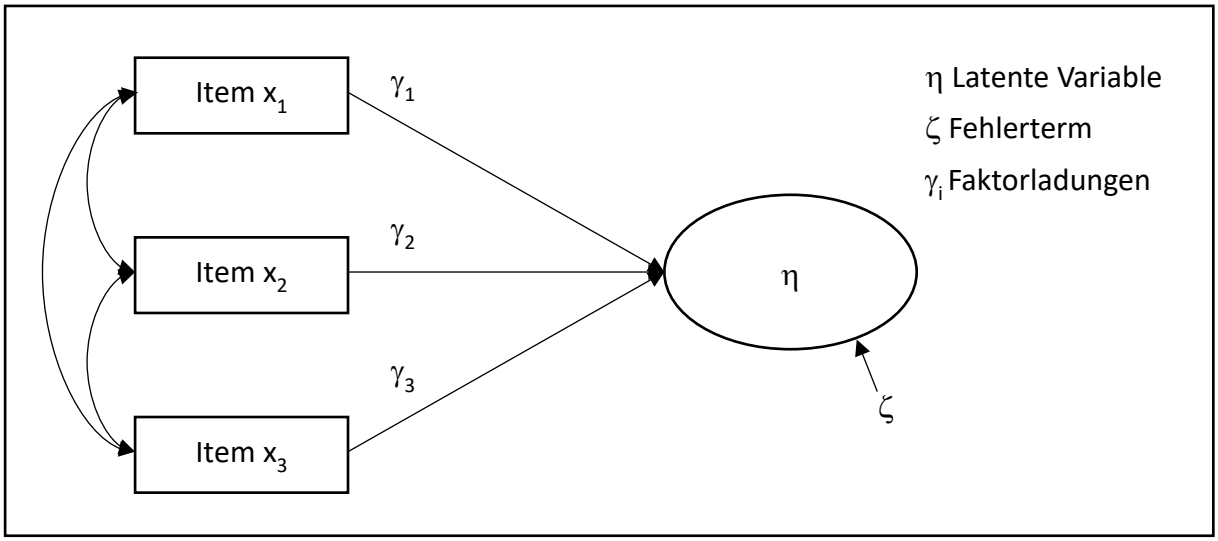


Abbildung 11: Formatives Messmodell (eigene Darstellung in Anlehnung an Edwards, 2011, S. 372)

Zwar können die Indikatorvariablen auch in einem formativen Messmodell miteinander korrelieren, eine Korrelation formativer Indikatoren ist jedoch keine notwendige Voraussetzung. Der zugrundeliegenden Kausalitätslogik zufolge hängt von den einzelnen Indikatoren jeweils nur das Konstrukt ab (Eberl, 2004; Edwards, 2011). In einem formativen Modell ist es daher möglich, dass sich eine Veränderung in der latenten Variablen nur in der Veränderung eines Indikators ablesen lässt (Eberl, 2004).

Im Gegensatz zur reflektiven Modellierung können dabei nicht ohne weiteres einzelne Indikatoren ohne Validitätsverlust ausgetauscht werden, da sich die Bedeutung des latenten Konstrukts aus den einzelnen Indikatoren zusammensetzt und ein Austausch oder Wegfall einer Messung den Charakter des Konstrukts verändern würde (Eberl, 2004; Kline, 2012). Ein klassisches Beispiel eines formativen Konstrukts ist der sozioökonomische Status (SES) einer Person, der sich als eine Art Sammelwert aus mehreren Indikatoren wie Einkommen, Schulbildung, Ausbildung und Wohnverhältnissen zusammensetzt (Fluck, 2020b; Urban & Mayerl, 2014).

Unterschiede in den Spezifikationsarten und Auswahl der Spezifikationsart für diese Arbeit

Neben dem fundamentalsten Unterschied, der im Vergleich zu reflektiven Modellen umgekehrten Kausalitätsrichtung formativer Messmodelle, bestehen auch bei der Beurteilung der Konstruktvalidität Unterschiede zwischen den Spezifikationsarten. Die herkömmlichen Kriterien zur Beurteilung der Konstruktvalidität wie die Höhe der Faktorladungen und Korrelationen zwischen den Indikatoren, die üblicherweise zur Bewertung reflektiver Modelle herangezogen werden, sind für formative Modelle i. d. R. nicht geeignet, weshalb auf alternative Methoden wie Expertenurteile und Kollinearitätsanalysen zurückgegriffen werden muss (Fluck, 2020b). Edwards (2011) beschreibt ausführlich weitere Unterschiede reflektiver und formativer Modellspezifikationen und legt dabei die Probleme formativer Modellierungen offen:

Zunächst unterscheiden sich reflektive und formative Messungen im Hinblick auf deren Dimensionalität. Reflektive Indikatoren repräsentieren i. d. R. eine Dimension und messen entsprechend ein zugrundeliegendes Konstrukt. Dies führt dazu, dass die einzelnen Indikatoren theoretisch austauschbar sind und das Entfernen eines Indikators die Bedeutung und Interpretation des Konstrukts nicht verändern sollte. Formative Indikatoren dagegen beschreiben jeweils verschiedene Dimensionen bzw. Facetten eines Konstrukts, weshalb häufig von Mehrdimensionalität formativer Messmodelle gesprochen wird. Jeder Indikator sollte im besten Fall einen Aspekt des Konstrukts beschreiben und die Redundanz der einzelnen Messungen sollte weitestgehend ausgeschlossen werden. Daher können nicht einfach Indikatoren entfernt werden, ohne den Charakter des Konstrukts zu verändern. Edwards (2011) wirft in diesem Kontext die Frage auf, ob ein solches Konstrukt, das durch konzeptuell heterogene Messungen repräsentiert wird, überhaupt nützlich ist oder ob dessen Mehrdeutigkeit einer sinnvollen Interpretation eventuell im Wege steht.

Weitere Unterschiede zwischen den Spezifikationsarten betreffen die interne Konsistenz. Da bei formativen Modellierungen keine (hohen) Korrelationen zwischen den Indikatoren vorausgesetzt werden, weil nach einer formativen Logik jeder Indikator eine einzigartige Facette erfasst, weisen formative Skalen i. d. R. eine mangelnde interne Konsistenz auf. Daraus können sich bei der Entscheidung für oder gegen eine der beiden Spezifikationsarten verschiedene Fehlschlüsse ergeben: Bspw. kommen Forscher*innen aufgrund der geringen internen Konsistenz einer Skala zu dem Schluss, dass die Indikatoren eher formativen als reflektiven Charakter haben, dabei könnte eine geringe interne Konsistenz auch auf eine schlecht konstruierte reflektive

Skala zurückzuführen sein. Generell ist eine geringe interne Konsistenz weder eine notwendige noch eine hinreichende Voraussetzung, um auf eine formative Messung zu schließen.

Auch hinsichtlich der Identifizierbarkeit der Messmodelle ergeben sich Unterschiede zwischen reflektiven und formativen Messungen. Während ein reflektives Konstrukt mit drei Indikatoren gerade identifiziert ist, ist ein formatives Modell immer unteridentifiziert, unabhängig von der Anzahl der Indikatoren. Dies erfordert die zusätzliche Spezifikation mindestens zweier reflektiver Messungen bzw. Indikatoren, die direkt oder indirekt durch die latente Variable beeinflusst werden, bspw. durch ein sog. *MIMIC (multiple indicator multiple cause) Modell*. Dies führt jedoch u. U. dazu, dass reflektive Indikatoren in ein Modell mit einbezogen werden (müssen), unabhängig davon, ob diese Messungen und deren Outcomes konzeptionell gerechtfertigt oder inhaltlich zielführend sind. Zusätzlich haben die reflektiven Indikatoren Einfluss auf die Ladungen der Indikatoren und beeinflussen dadurch wiederum die Bedeutung des Konstrukts.

Ein weiterer bedeutender Unterschied betrifft die Spezifikation von Messfehlern. Während reflektive Indikatoren jeweils einen individuellen Fehlerterm besitzen, wird dieser in formativen Modellen dem latenten Konstrukt zugeordnet. Dieser Fehlerterm sollte jedoch nicht wie im reflektiven Modell als Messfehler interpretiert werden, sondern als Aspekt des Konstrukts, welcher nicht durch die Indikatoren erklärt wird. Die Indikatormessungen sind nach dem Verständnis formativer Modellierung demnach frei von (Mess-)Fehlern. Diese Annahme einer fehlerfreien Messung ist jedoch schwer damit zu vereinbaren, dass die Messungen i. d. R. mit bestimmten sozialwissenschaftlichen Methoden erhoben wurden, die per se nicht fehlerfrei messen können, da es bspw. zu Abweichungen in der Item-Interpretation, Kodierfehlern, technischen Problemen oder sonstigen Varianzen im Messprozess kommen kann. Die Annahme einer fehlerfreien Indikatormessung eines formativen Modells konterkariert somit einen der Hauptvorteile von Strukturgleichungsmodellen, die explizite Berücksichtigung von Messfehlern bei der Modellierung latenter Konstrukte.

Auch in der empirischen Anwendung erweisen sich formative Modellierungen als nicht unproblematisch. An dieser Stelle sei auf die Habilitationsschrift von Fluck (2020a) verwiesen, die sich ausführlich u. a. mit der empirischen Bestimmung reflektiver bzw. formativer Spezifikationsarten beschäftigt. Die Autorin beschreibt in ihrer Arbeit ausführlich ein Auswahlschema zur Bestimmung der korrekten Spezifikation und unterzieht dieses einer empirischen Prüfung. Die Ergebnisse zeigen, dass eine Entscheidung für bzw. gegen die eine oder andere Spezifikationsart häufig alles andere als eindeutig zu treffen ist. Selbst eine strukturierte Überprüfung

mittels TETRAD-Test oder CFA lässt teils keine eindeutige Entscheidung bzw. Ablehnung einer der beiden Modellierungsvarianten zu. Auch wenn aufgrund anfänglicher Überlegungen eine formative Spezifikation gewählt wird, ist nicht gesichert, ob sich diese in der Modellierung bewährt.

Generell besteht die Kritik, dass formative Messmodelle eher schwer zu rechnen und daher häufig wenig praktikabel sind (Fluck, 2020b). Mit den gängigen Statistikprogrammen, in denen i. d. R. varianzbasierte Verfahren zur Modellschätzung implementiert sind, lassen sich formative Modellierungen nicht ohne weiteres realisieren. Eine geeignete Alternative zur Schätzung formativer Modelle stellt die kovarianzbasierte *Partial Least Squares (PLS)-Methode* dar (Fluck, 2020a). Das PLS-Verfahren weist jedoch andererseits einige Nachteile auf, wie den Mangel an globalen und lokalen inferenzstatistischen Gütemaßen. Zusätzlich ist der PLS-Ansatz eher als hypothesengenerierender Ansatz zu verstehen und ist weniger geeignet zum Testen von Theorien (Eberl, 2004; Fluck, 2020a), was jedoch ein Ziel dieser Arbeit darstellt.

In vorliegender Arbeit wird auf eine reflektive Modellierung der untersuchten latenten Konstrukte zurückgegriffen, was dem Vorgehen der meisten Arbeiten zum TAM entspricht, in deren Forschungskontext die Ergebnisse dieser Untersuchung letztendlich zu verorten sind. Aus Gründen der Praktikabilität wird auf eine detaillierte Prüfung der einzelnen Voraussetzungen der einen oder anderen Modellierungsart verzichtet, weil dies den Rahmen dieser Arbeit sprengen würde und ggf. dennoch, aus den aufgeführten Gründen, nicht zu einer eindeutigen Entscheidung führen würde. Zudem trägt die Entscheidung für eine reflektive Modellierung einem der zentralen Vorteile von Strukturgleichungsmodellen Rechnung, der Berücksichtigung von Messfehlern, was im Falle einer formativen Modellierung nicht möglich wäre.

Korrelierte Fehlerterme

Im Hinblick auf die Fehlerterme der Indikatorvariablen reflektiver Messmodelle ergibt sich eine weitere Besonderheit der Modellierung latenter Konstrukte und Strukturgleichungsmodelle, die in diesem Abschnitt beschrieben wird. Strukturgleichungsmodelle ermöglichen die Kontrolle systematischer Verzerrungen durch Messfehler, die zusätzlich zu den zufälligen Messfehlern auftreten können, indem Kovarianzstrukturen zwischen Messfehlern modelliert werden, bspw. durch die Spezifikation von Korrelationen bestimmter Residuen (Bollen, 2014; Urban & Mayerl, 2014). Ein Zufallsfehler ist die zufällige Abweichung einer Messung vom theoretisch wahren Wert, der die Reliabilität einer Messung beeinflusst, während systematische Fehler die

Validität einer Messung mitbestimmen (Söhnchen, 2009). Für derartige systematische Fehler, die sich in einer substanziell bedeutsamen Fehlerkorrelation niederschlagen können, gibt es verschiedene mögliche Ursachen. Eine Erklärung könnte ein weiterer, im Modell nicht berücksichtigter Faktor sein, der auf zwei oder mehr Indikatoren wirkt. In Längsschnittanalysen kommt es typischerweise häufig zu korrelierten Fehlern zwischen identischen Indikatoren, die zu verschiedenen Messzeitpunkten erhoben wurden. Eine der häufigsten Quellen korrelierter Fehler bilden des Weiteren Methodeneffekte (Schermelleh-Engel & Schweizer, 2012; Urban & Mayerl, 2014). Hierbei zeigt sich ein ebenfalls im Modell unberücksichtigter Methodenfaktor als verantwortlich für eine gemeinsame Fehlervarianz bestimmter Items. Dieser entsteht bspw. durch ähnlich formulierte Fragentexte oder sonstige Spezifika eines für alle betroffenen Messungen eingesetzten Erhebungsinstruments (Urban & Mayerl, 2014). Dieser Annahme, dass den Ausprägungen von zwei oder mehreren Indikatorvariablen andere Ursachen zugrunde liegen als der gemeinsame Einfluss des latenten Faktors, kann durch die bereits erwähnte Spezifikation korrelierter Residuen oder Methodenfaktoren der betroffenen Items begegnet werden (Brown & Moore, 2012).

Konkret bedeutet die Spezifikation korrelierter Residuen, dass der Fehlerterm einer Variablen mit dem Fehlerterm einer anderen Variablen korreliert ist. Korrelierte Fehlerterme sind sowohl innerhalb der Indikatorvariablen eines Konstrukts denkbar, als auch Konstrukt übergreifend (Rubio & Gillespie, 1995). Visualisiert werden diese *Residualkorrelationen* durch ungerichtete Pfeile zwischen den betroffenen Messfehlern. In dem in Abbildung 9 dargestellten Strukturmodell repräsentieren die ungerichteten Pfeile zwischen δ_1 und δ_2 sowie zwischen δ_2 und δ_3 die Residualkorrelation zwischen den Indikatoritems x_1 und x_2 bzw. x_2 und x_3.

Üblicherweise werden die *Residualvarianzen* in einem Messmodell als unkorreliert angenommen und die Korrelationen nicht zur Schätzung freigegeben, jedoch ist dieses Axiom in der Praxis nicht immer haltbar (Berning, 2019; Westfall, Henning & Howell, 2012). Da nicht berücksichtigte Messfehlerkorrelationen in einem Modell zu verzerrten Schätzungen und einem mangelnden Modellfit führen können, empfehlen Rubio und Gillespie (1995) ein Modell auf korrelierte Fehler zu untersuchen und diese bei vorliegender Signifikanz in die Modellspezifikation aufzunehmen. Andere Autor*innen argumentieren jedoch, dass auch die Modellierung von Messfehlerkorrelationen zu Verzerrungen in der Parameterschätzung führen kann und Fehlerkorrelationen häufig aus den falschen Gründen, bspw. zur reinen Verbesserung des Modellfits, spezifiziert werden (Hermida, 2015).

Trotz aller Kritik muss in der Praxis i. d. R. eine Lösung zur Modellierung gemeinsamer Fehlerursachen wie potenzielle Methodenfaktoren gefunden werden. Da nicht beliebig viele Faktoren in ein (sparsames) Modell miteinbezogen werden können, bieten korrelierte Messfehler eine pragmatische und praktikable Lösung (Westfall et al., 2012), die auch in dieser Arbeit angewandt wird. Dabei gilt jedoch in jedem Fall zu beachten, dass die zugelassenen Residualkorrelationen immer plausibel theoretisch oder empirisch begründbar sind und niemals ausschließlich zur Verbesserung des Modellfits, bspw. basierend auf Modifikationsindizes, spezifiziert werden (Boomsma, 2000; Rubio & Gillespie, 1995; Urban & Mayerl, 2014). Des Weiteren sollte bei der Spezifikation korrelierter Messfehler konsistent vorgegangen werden. Wenn eine gemeinsame Ursache für die Ausprägung von Messfehlern bei bestimmten Indikatoren unterstellt wird, sollte eine Korrelation auch für alle betroffenen Residuen spezifiziert werden (Brown & Moore, 2012).

4.3.3. Vorgehensweise bei Strukturgleichungsanalysen

Die Anwendung von Strukturgleichungsanalysen folgt in der Forschungspraxis häufig den Arbeitsschritten, die im folgenden Kapitel erläutert werden (siehe bspw. A. Fuchs, 2011; Kline, 2011; Ullman & Bentler, 2003):

Schritt 1: Modellspezifikation

Dieser erste Arbeitsschritt erfordert eine umfassende theoretische Vorarbeit, auf deren Basis ein zu testendes Modell entwickelt wird und die entsprechenden Hypothesen über spezifizierte Wirkungszusammenhänge aufgestellt werden. Die hypothetischen Zusammenhänge werden hierbei mittels Pfadmodellen visualisiert (siehe Kapitel 3) (A. Fuchs, 2011; Kline, 2011; Ullman & Bentler, 2003). Auch mögliche spätere Modellanpassungen, die aus theoretischer Perspektive oder basierend auf empirischen Erkenntnissen gerechtfertigt wären, können in diesem Schritt bereits mitbedacht werden, um bei einer möglichen nachträglichen Modellanpassung den gleichen theoretischen Ansprüchen gerecht zu werden wie bei der Spezifizierung des ursprünglichen Modells (siehe auch Schritt 6) (Kline, 2011). Der Modellformulierung folgt i. d. R. die Operationalisierung der zu untersuchenden Konstrukte (siehe Kapitel 4.1.1), auf deren Basis die Modellschätzung erfolgen kann (A. Fuchs, 2011; Kline, 2011).

Schritt 2: Prüfung auf Identifizierbarkeit

Die Identifizierbarkeit eines Modells ist die wichtigste und erste Voraussetzung für eine erfolgreiche Modellschätzung. Ein Modell ist dann identifizierbar, wenn es, basierend auf der Anzahl der verfügbaren empirischen Informationen, theoretisch möglich ist, eine Schätzung aller zu schätzenden Modellparameter zu erhalten (Kline, 2011; Urban & Mayerl, 2014). Identifizierbarkeit bezieht sich somit auf das Verhältnis bekannter und unbekannter Informationen innerhalb eines Modells und die Frage, ob genug bekannte Informationen vorliegen, um alle spezifizierten unbekannten Informationen zu schätzen und eine Lösung für das aufgestellte Gleichungssystem zu erhalten. Die Anzahl bekannter Informationen innerhalb eines SEM ergibt sich i. d. R. aus der Anzahl der Elemente der beobachteten Varianz-Kovarianz-Matrix. Ein Modell mit k beobachteten Variablen verfügt über k(k+1)/2 bekannte Werte. Die unbekannten Informationen eines spezifizierten Modells umfassen alle zu schätzenden Parameter wie Varianzen, Kovarianzen, Faktorladungen und Strukturkoeffizienten (Kenny & Milan, 2012). Bei der Beurteilung der Identifizierbarkeit sind u. a. die Freiheitsgrade (df - *degrees of freedom*) des Modells zu beachten, welche die Differenz zwischen bekannten und unbekannten Informationen widerspiegeln. Übersteigt die Anzahl unbekannter Parameter die Anzahl bekannter Informationen ($df_M < 0$) und gibt es keine eindeutige Lösung für die Modellgleichung, gilt ein Modell als unteridentifiziert. Entspricht die Anzahl freier Parameter derjenigen der empirischen Beobachtungen, besitzt ein Modell also null Freiheitsgrade ($df_M = 0$), ist dieses gerade identifiziert bzw. saturiert. Übersteigt die Anzahl der verfügbaren Informationen die der zu schätzenden Parameter ($df_M > 0$), handelt es sich um ein überidentifiziertes Modell (Kenny & Milan, 2012; Kline, 2011).

Bei der Spezifizierung von Strukturgleichungsmodellen sollte immer ein überidentifiziertes Modell angestrebt werden. Ein unteridentifiziertes Modell ist nicht lösbar, ein saturiertes Modell fittet dagegen immer perfekt, da es nur eine Lösung für das Modellgleichungssystem gibt, und lässt somit keine Falsifizierbarkeit zu. Ein Modell sollte jedoch immer falsifizierbar sein, da nur unter dieser Voraussetzung ein Modell gefunden werden kann, welches möglichst gut zu den vorhandenen empirischen Daten passt. Ein gerade identifiziertes Modell ist jedoch weder falsifizierbar noch sind die Fitindikatoren (sinnvoll) interpretierbar. Ein überidentifiziertes Modell hingegen passt nie perfekt zu den Daten und ermöglicht dadurch die Betrachtung des Grades der Abweichung anhand entsprechender Fitindikatoren sowie die Beurteilung der Modellpassung zu den Daten (Kenny & Milan, 2012). Bei unter- oder gerade identifizierten Modellen ist es möglich, bspw. durch die Fixierung oder Gleichsetzung bestimmter Parameter,

bestimmte Restriktionen zu definieren, um mehr Freiheitsgrade zu schaffen und ein identifiziertes Modell zu erhalten (Gana & Broc, 2019). Eine pragmatische Möglichkeit zur Überprüfung der Identifizierbarkeit ist die Schätzung des spezifizierten Modells mithilfe der gewählten Software. Wenn das Modell ohne Fehlermeldungen geschätzt wird, muss es identifiziert sein. Zwar gibt es elaboriertere Methoden zur Überprüfung der Identifizierbarkeit, deren Ausführung würde jedoch für diese Arbeit keinen Mehrwert liefern (Kenny & Milan, 2012). Die Prüfung der Identifizierbarkeit findet somit praktisch erst in Schritt 4 zusammen mit der Modellschätzung statt.

Schritt 3: Empirische Datenerhebung und Aufbereitung

Gilt ein aufgestelltes Modell als theoretisch identifizierbar, werden Daten gesammelt, anhand derer das Modell getestet werden kann, und entsprechend für die weitere Nutzung zur Modellschätzung aufbereitet (Kline, 2011). Die entsprechenden Arbeitsschritte sind für diese Arbeit in den Kapiteln 4.1.2 und 4.2 beschrieben.

Ein Aspekt, der bei der Datenerhebung grundsätzlich bedacht werden sollte, ist die geplante Größe der Stichprobe, da Strukturgleichungsmodelle je nach Komplexität eine bestimmte Stichprobengröße voraussetzen. Eine zu kleine Stichprobe kann sich, wie jedoch auch eine zu große (siehe Abschnitt zu Chi-Quadrat.), u. a. auf Teststatistiken auswirken, ebenso wie auf die statistische Power. Zwar gibt es viele Arbeiten zu diesem Thema, eindeutige Empfehlungen oder Regeln hinsichtlich einer optimalen Fallzahl auszumachen, ist jedoch nicht trivial (Urban & Mayerl, 2014). Jackson (2003) plädiert für die N:q Regel, die die Anzahl der Fälle ins Verhältnis zur Anzahl der frei zu schätzenden Parameter setzt, und ein optimales Verhältnis von 20:1, also 20 Fälle je freiem Parameter, festlegt. Andere Ansätze fokussieren eher auf die absolute Größe einer Stichprobe und betrachten die typische Stichprobengröße von N = 200 in SEM-Untersuchungen als Referenz für eine angemessene Stichprobengröße (Kline, 2011) oder verlangen, je nach Schätzmethode, Stichprobengrößen zwischen 500 und 2500 (Ullman & Bentler, 2003).

Schritt 4: Modellschätzung

Mit der Modellschätzung erfolgt die empirische Überprüfung der Passung des theoretischen Modells zu den erhobenen Daten (Kline, 2011). Das Ziel einer Modellschätzung ist die Minimierung der Differenz zwischen empirischer Varianz-Kovarianz-Matrix und geschätzter

Kovarianzmatrix. Je kleiner die Differenz und je größer somit die Übereinstimmung zwischen beobachteter und geschätzter Varianz-Kovarianz-Matrix, desto besser passt ein Modell zu den Daten (Reinecke, 2014; Ullman & Bentler, 2003). Zur Beurteilung dieser Modellpassung stehen verschiedene Maße zur Verfügung, die unter Schritt 5 im Detail erläutert werden.

Bevor eine Modellschätzung durchgeführt werden kann, müssen jedoch mehrere Entscheidungen getroffen werden: Das Vorgehen bei der Modellschätzung sollte festgelegt und ein passendes Schätzverfahren ausgewählt werden. Zusätzlich muss über den Umgang mit fehlenden Werten bei der Modellschätzung entschieden und eine Auswertungssoftware bestimmt werden. Diese Schritte, deren Reihenfolge durchaus variabel ist, sind im Folgenden beschrieben.

Strategie der Modellschätzung.

Zur Modellschätzung bei Strukturgleichungsanalysen werden in der Literatur verschiedene Ansätze erläutert. Urban und Mayerl (2014) schlagen hierbei in Anlehnung an Bollen (2000) u. a. folgende Strategien vor: (1) *Four Step*-Strategie, (2) *Two Step*-Strategie, (3) *One Step*-Strategie, (4) *separate factor analysis*-Strategie. Die ersten beiden Varianten (Four Step- und Two Step-Strategie) basieren, ebenso wie der separate factor analysis-Ansatz, in unterschiedlicher Graduierung auf getrennten Schätzungen von Messmodellen und Strukturmodell, während gemäß der One Step-Strategie von Beginn an simultan sowohl Messmodelle als auch kausale Struktur zwischen den latenten Variablen getestet werden. Auch wenn der einschrittige Ansatz der Grundidee von SEM einer simultanen Schätzung von Messmodellen und Strukturbeziehungen am meisten entspricht, ist dieser jedoch in der Umsetzung sehr komplex und kann häufig zu invaliden Modellanalysen führen. Dabei ist es schwierig, die Ursachen für nicht fittende Modelle eindeutig auf Probleme der Struktur- oder Messmodellebene zurückzuführen (Bollen, 2000; Reinecke, 2014; Urban & Mayerl, 2014). Da dieser Ansatz zusätzlich besonders für die Analyse inhaltlich-theoretisch stark fundierter Struktur- und Messmodelle geeignet ist (Reinecke, 2014), während in dieser Arbeit teils neu konstruierte Skalen verwendet werden, wurde eine mehrschrittige Analysestrategie gewählt.

Besonders populär ist unter diesen der Two Step-Ansatz, demnach in einem ersten Schritt mittels CFA simultan die Messmodelle aller latenten Variablen geschätzt werden, wobei alle latenten Variablen kovariieren dürfen. Ist die Schätzung der CFA zufriedenstellend, werden im zweiten Schritt die Kovarianzen durch die postulierte Kausalstruktur des Strukturmodells ersetzt (Anderson & Gerbing, 1988; Steinmetz, 2015). Dieses Vorgehen ermöglicht eine separate Fehleranalyse bzw. -beseitigung bei Fehlspezifikationen zunächst auf Messmodellebene und

dann auf Strukturebene (Reinecke, 2014). Noch kleinschrittiger ist das Vorgehen der separate factor analysis-Strategie, bei welcher jedes Messmodell zunächst einzeln getestet wird, bevor passende Messmodelle in ein Strukturmodell aufgenommen werden (Urban & Mayerl, 2014). Bei diesem Vorgehen besteht die Kritik, dass einzelne Messmodelle durchaus gut fitten können, jedoch die Gefahr einer falschen Indikatorzuordnung zu einem Faktor besteht, weil die Faktoren separat geschätzt werden (Bollen, 2000; Urban & Mayerl, 2014).

Trotz dieser Kritik wurde in dieser Arbeit aus Gründen der Praktikabilität eine separate factor analysis-Strategie gewählt, da die theoretische Modellstruktur sehr komplex ist, sodass eine simultane Testung aller latenten Konstrukte, wie bspw. in dem zweischrittigen Ansatz empfohlen, nicht trivial und dafür tendenziell fehleranfällig wäre. Um möglichen Missspezifikationen im Sinne fehlerhafter Itemzuordnungen zu begegnen, wurden jedoch in einem weiteren Analyseschritt zusätzlich alle latenten Konstrukte in einer gemeinsamen CFA geschätzt. Nach erfolgreicher Schätzung der Messmodelle wurden jeweils die Strukturmodelle analysiert.

Auswertungssoftware.

Die statistischen Auswertungen dieser Arbeit sowie sämtliche Vorarbeiten zur Datenaufbereitung wurden mit Hilfe der freien syntaxbasierten Programmiersprache R (Version 3.5.1) durchgeführt (R Core Team, 2018). Für Datensatzaufbereitung und deskriptive Analysen wurden dabei die Pakete dplyr, Version 0.8.5 (Wickham, Francois, Henry & Müller, 2020), tidyr, Version 1.0.0 (Wickham & Henry, 2019), knitr, Version 1.3 (Xie, 2021), ggplot2, Version 3.3.3 (Wickham, 2016) und psych, Version 1.8.12 (Revelle, 2018) eingesetzt. Das Paket car (Version 3.0-10) wurde u. a. für die Berechnung von Levene Tests genutzt (Fox & Weisberg, 2019), das Paket semTools (Version 0.5-4) (Jorgensen, Pornprasertmanit, Schoemann & Rosseel, 2021) u. a. zur Berechnung der durchschnittlich erfassten Varianz. Zur Datenvisualisierung wurde das Paket ggplot2 (Version 3.3.3) (Wickham, 2016) verwendet. Für die zentralen statistischen Auswertungen, der Schätzung von CFAs und Strukturgleichungsmodellen, wurde das Paket lavaan (Version 0.6-7) eingesetzt (Rosseel, 2012).

Auswahl des Schätzverfahrens.

Die Qualität und das Ergebnis einer Modellschätzung einer Strukturgleichungsanalyse hängt immer auch von der Auswahl eines passenden Schätzverfahrens ab (Fan, Thompson & Wang, 1999; Lei & Wu, 2012). In der Forschung haben sich hierfür einerseits der Ansatz eines varianzbasierten Verfahrens (PLS - partial least squares) und andererseits kovarianzbasierte

Verfahren etabliert (A. Fuchs, 2011)[5]. Bei den kovarianzbasierten Verfahren, die auch in dieser Arbeit zum Einsatz kommen, erfolgt jede Modellschätzung durch die Minimierung der Diskrepanz einer empirischen Varianz-Kovarianz-Matrix und einer Modell-implizierten Varianz-Kovarianz-Matrix, bei welcher die unbekannten Modellparameter mittels eines geeigneten Schätzverfahrens geschätzt werden. Die Modellschätzung folgt einem iterativen Prozess, bei dem Startwerte für die zu schätzenden Parameter festgelegt werden (üblicherweise durch die genutzte SEM-Software), aus welchen eine erste vorläufige Varianz-Kovarianz-Matrix berechnet wird. Durch die Minimierung der Differenz zwischen dieser vorläufigen Matrix und der empirischen Varianz-Kovarianz-Matrix erfolgen neue Parameterschätzungen, die die alten ersetzen. Dieser Prozess wiederholt sich so lange, bis die Anpassungen der Parameterschätzungen zwischen zwei Iterationen nur noch zu marginalen Reduktionen der Differenz zwischen empirischer und reproduzierter Varianz-Kovarianz-Matrix führen, und somit das Konvergenzkriterium erreicht ist. Sobald die Veränderungen also geringer sind als dieser gesetzte Grenzwert, wird der iterative Prozess abgebrochen und die letzten Parameterschätzungen als finale Lösung ausgegeben (Gana & Broc, 2019; Lei & Wu, 2012).

Für den beschriebenen Prozess der Modellschätzung stehen im Bereich der kovarianzbasierten Verfahren verschiedene *Diskrepanzfunktionen* bzw. Schätzmethoden wie die *Maximum-Likelihood (ML), Unweighted-Least-Square (ULS),* die *Generalized-Least-Square (GLS)* oder die *Weighted-Least-Square (WLS) Methode* zur Verfügung (Reinecke, 2014). Die Methoden unterscheiden sich hinsichtlich der zu minimierenden Diskrepanzfunktionen und der daraus berechneten χ^2-Statistik zur Beurteilung des Modellfits (Gana & Broc, 2019; Lei & Wu, 2012). Das prominenteste Verfahren der Modellschätzung, das auch in den meisten Softwarepaketen als Default Einstellung implementiert ist, ist die Maximum-Likelihood-Methode (ML) (Beaujean, 2014; Keith, 2019; Kline, 2011; Schermelleh-Engel, Moosbrugger & Müller, 2003)[6]. Die ML-Methode zielt darauf ab, Schätzungen zu erzeugen, die mit der höchsten Wahrscheinlichkeit die empirische Stichprobe reproduzieren. Dabei wird für jedes Set möglicher Parameter die Wahrscheinlichkeit berechnet, mit der die Schätzwerte die empirischen Daten abbilden. Die Schätzungen mit der höchsten Wahrscheinlichkeit werden ausgegeben (Keith, 2019). Aus dem in der ML-Schätzung ermittelten Funktionswert lässt sich die χ^2-Verteilung berechnen, anhand

[5] Da in dieser Arbeit mit der robusten ML-Methode ein kovarianzbasiertes Verfahren genutzt wird, werden nur diese im Folgenden näher erläutert. Für weitere Informationen zum PLS-Schätzverfahren sei bspw. auf Hair et al. (2017) und Monecke und Leisch (2012) verwiesen.

[6] Die weiteren Schätzverfahren werden hier nicht weiter erläutert, für weiterführende Informationen siehe bspw. Reinecke (2014).

derer die Modellpassung beurteilt wird. Je größer der χ^2-Wert (und je kleiner der p-Wert), desto eher liegt eine Diskrepanz zwischen Modell und Daten vor (Reinecke, 2014).

Zwar handelt es sich bei der ML-Methode um einen sehr performanten, jedoch auch um einen relativ voraussetzungsvollen Ansatz, der u. a. ein metrisches Messniveau, einen großen Stichprobenumfang und insbesondere eine multivariate Normalverteilung fordert (Beaujean, 2014). Die Voraussetzung eines metrischen Skalenniveaus kann, trotz des eigentlich ordinalen Charakters der in Befragungsstudien und auch in dieser Arbeit verwendeten Likert Skala, als gegeben angenommen werden. Eine Likert skalierte Variable kann demnach ähnlich zu einer Intervallskala betrachtet werden, sofern Äquidistanz und Symmetrie der Skala gegeben sind, und entsprechend für statistische Verfahren, wie bspw. Strukturgleichungsmodellierungen, genutzt werden (Hair et al., 2017). Die Voraussetzung einer multivariaten Normalverteilung der manifesten Variablen ist hingegen i. d. R. in sozialwissenschaftlichen Untersuchungen kaum zu erfüllen (Beaujean, 2014; Reinecke, 2014). Die ML-Schätzmethode erweist sich zwar im Hinblick auf die Parameterschätzung als durchaus robust gegen eine Verletzung der Annahme einer multivariaten Normalverteilung, jedoch ist die Schätzung bei nicht normalverteilten Daten nur bedingt effizient (Schermelleh-Engel et al., 2003).

Da bei der in dieser Arbeit genutzten Datengrundlage nicht von einer multivariaten Normalverteilung ausgegangen werden kann (siehe Kapitel 5.1.1) und eine ML-Schätzung des Weiteren nicht für Modelle mit korrelierten Residuen geeignet ist (Rubio & Gillespie, 1995), insbesondere nicht bei Residualkorrelationen zwischen exogenen und endogenen Variablen (Urban & Mayerl, 2014), wurde für diese Arbeit mit der *robusten Maximum-Likelihood (MLR) Schätzung* eine angepasste Variante dieses Verfahrens gewählt. Hierbei handelt es sich um ein robustes ML-Verfahren, welches durch die Berücksichtigung der Abweichungen von der multivariaten Normalverteilung verteilungsrobuste Standardfehler und χ^2-Statistiken berechnet und entsprechend korrigierte Schätzungen liefert (Reinecke, 2014; Urban & Mayerl, 2014). Die Korrektur der Test-Statistik erfolgt gemäß dem Ansatz von Yuan und Bentler (1998, 2000). Berechnet wird die Yuan-Bentler (YB)-χ^2-Statistik (Reinecke, 2014) sowie der robuste (Hubert-White) Standardfehler (Beaujean, 2014). Weitere Informationen zur (angepassten) χ^2-Statistik und den anderen (robusten) Fitstatistiken liefert der Abschnitt zur Modellevaluation (siehe S. 149ff.).

Umgang mit fehlenden Werten – FIML-Methode.

Im Zusammenhang mit der Auswahl eines geeigneten Schätzverfahrens steht i. d. R. auch die Entscheidung über den Umgang mit fehlenden Werten, weil eine Datenbereinigung meist – wie auch in dieser Arbeit – nicht zu ausschließlich vollständigen Fällen führt.

Hierbei stellt sich zunächst die Frage nach dem zugrundliegenden Ausfallmechanismus bzw. der Ursache für die fehlenden Werte, da dieser bei der späteren Datenanalyse Einfluss auf das Verfahren zur Behandlung fehlender Daten hat. Fehlende Daten werden dabei entweder als MCAR (*missing completely at random*), MAR (*missing at random*) oder MNAR (*missing not at random*) charakterisiert. Tritt ein Ausfall rein zufällig auf und hängen die fehlenden Werte einer Variablen nicht mit dieser oder einer anderen Variablen zusammen, liegt MCAR vor. Von MAR wird gesprochen, wenn fehlende Werte einer Variablen unabhängig von der Variablen selbst sind, jedoch abhängig von anderen Variablen des Datensatzes, also durch andere Variablen erklärt werden können. Entgegen der Bezeichnung, die eine Zufälligkeit impliziert, wird diesem Ausfallmechanismus somit eine systematische Beziehung unterstellt. Treten fehlende Werte einer Variablen nicht zufällig auf, sondern werden von der betroffenen Variablen und anderen Variablen beeinflusst, spricht man von MNAR. Werden Daten, denen dieser Ausfallmechanismus zugrunde liegt, für statistische Analysen genutzt, kann dies zu einem Bias in den Ergebnissen führen (Enders, 2010; Graham et al., 2003; Graham, 2012; Little & Rubin, 2002; Schafer & Graham, 2002).

Zum statistischen Umgang mit fehlenden Werten stehen daher verschiedene Verfahren zur Verfügung.[7] Während klassische Methoden wie listen- oder paarweiser Fallausschluss zu Daten- bzw. Informationsverlust und einer Verkleinerung der Stichprobe führen (Gana & Broc, 2019; Schafer & Graham, 2002), wurden mittlerweile in den meisten Programmen zur SEM-Schätzung deutlich elaboriertere Methoden zum Umgang mit fehlenden Werten, wie die *full information estimation maximum likelihood* (FIML) Schätzung, implementiert (Brown, 2015; Keith, 2019). Bei Anwendung dieser Methode werden fehlende Werte in einem Schritt zusammen mit der Parameterschätzung und der Schätzung der Standardfehler behandelt (Graham, 2009).

Voraussetzung für die Nutzung der FIML-Methode ist mindestens die Annahme des Vorliegens von MAR der fehlenden Werte. In der Praxis ist das Vorliegen von MAR häufig nur eine Annahme, die nicht wirklich überprüfbar ist (Enders, 2010; Schafer & Graham, 2002), jedoch zeigt

[7] Für weitere Informationen zu verschiedenen Verfahren im Umgang mit fehlenden Werten siehe bspw. Enders (2010); Graham (2012); Little und Rubin (2002).

sich, dass selbst eine irrtümliche Unterstellung von MAR i. d. R. kaum einen Einfluss auf die Ergebnisse einer Modellschätzung hat (Schafer & Graham, 2002). Besonders bei großen Datensätzen mit prozentual geringem fehlenden Anteil von maximal 5 % schlagen sich der zugrundeliegende Ausfallmechanismus bzw. das daraus abgeleitete Verfahren zum Umgang mit den fehlenden Daten kaum in den Ergebnissen nieder (Tabachnick & Fidell, 2010). Aufgrund der Stichprobengröße und des insgesamt eher geringen Anteils fehlender Werte in den finalen Datensätzen (siehe auch Kapitel 5.1.1) und da zudem keine Hinweise explizit dagegensprechen, wird für die Datengrundlage dieser Arbeit MAR angenommen und die FIML-Methode angewendet.

Mit der FIML-Methode erfolgt die Modellschätzung im Grundsatz analog zur ML bzw. MLR-Schätzung. Der FIML-Algorithmus berechnet jeweils eine fallweise Likelihood-Funktion unter Einbezug aller für einen Fall verfügbaren Variablen. Diese fallweisen Funktionen werden über das gesamte Sample akkumuliert und maximiert. Dies führt dazu, dass alle verfügbaren Fälle trotz einzelner fehlender Daten genutzt werden, wobei keine Imputation der fehlenden Daten stattfindet (Enders & Bandalos, 2001). Dieser Ansatz liefert weitestgehend unverzerrte Parameterschätzung und Standardfehler (Enders, 2010). Zudem demonstrieren Simulationsstudien die Überlegenheit der FIML-Methode gegenüber herkömmlichen Verfahren (Enders & Bandalos, 2001), insbesondere im Falle nicht normalverteilter Daten (Enders, 2001). Aus rein praktischen Gründen ist die FIML-Methode ebenfalls zu bevorzugen, da sie als modellbasierter Ansatz das Problem der fehlenden Daten und die Parameterschätzung direkt in einem Schritt löst (Graham et al., 2003). Auch in lavaan ist die FIML-Methode zusammen mit einer MLR-Schätzung niedrigschwellig implementiert, wobei die Behandlung fehlender Werte direkt mit der Modellschätzung erfolgt (Beaujean, 2014) und im Gegensatz zu datenbasierten Verfahren wie bspw. multipler Imputation keinen zusätzlichen Arbeitsschritt durch vorheriges Erzeugen eines neuen vollständigen Datensatzes erfordert (Reinecke, 2014). Deshalb wurde auch in dieser Arbeit dem Problem unvollständiger Datensätze mit dem FIML-Ansatz begegnet.

Schritt 5: Modellevaluation – Relevante Gütekriterien

Zur Modellevaluation, also der Beurteilung der Passung eines geschätzten Modells (CFA oder SEM) und der Interpretation der Modellparameter, stehen verschiedene Kriterien zur Verfügung. Die für diese Arbeit relevanten Gütekriterien und deren Grenzwerte werden im Folgenden erläutert. Zusammenfassend sind diese außerdem in Tabelle 16 auf S. 153 dargestellt.

Der χ^2-Test.

Zunächst gilt es, die grundsätzliche Passung des Modells zu beurteilen, d. h. den Fit zwischen empirischer und geschätzter Kovarianzmatrix. Ein häufig verwendetes Gütekriterium ist der Chi-Quadrat-Test (auch *Likelihood-Ratio*-Statistik) (siehe auch Auswahl des Schätzverfahrens, S. 145ff.). Der Test zur Überprüfung des exakten Modellfits prüft die Nullhypothese der Passung der modelltheoretischen Kovarianzmatrix zu den wahren Werten der Grundgesamtheit gegen die entsprechende Alternativhypothese (Brown, 2015; Weiber & Mühlhaus, 2014). Ist der p-Wert der χ^2-Statistik > .05, kann die Nullhypothese nicht zurückgewiesen werden und das Modell gilt als geeignet, die empirische Kovarianzmatrix abzubilden (Schermelleh-Engel et al., 2003), wohingegen die Nullhypothese bei einem signifikanten χ^2-Wert zurückgewiesen werden muss (Brown, 2015).

Die Nullhypothese des χ^2-Tests der Gleichheit von empirischer und geschätzter Kovarianzmatrix ist jedoch eine sehr restriktive Annahme, die bei großen Stichproben mit großer Sicherheit zu einer Ablehnung führt (Finch & French, 2015). Hierin liegt eine grundsätzliche Limitation dieser Statistik, die sehr sensibel bei großen Stichproben ist, mitunter schon bei $N \geq 400$ (Berning, 2019). Auch kleine Abweichungen zwischen modellimplizierter und empirischer Kovarianzmatrix können bei großem N bereits zu einer Verwerfung der Nullhypothese und der Ablehnung eines plausiblen Modells führen (Schermelleh-Engel et al., 2003). Der mittels MLR-Schätzmethode berechnete robuste χ^2-Test (χ_r^2–Test) begegnet zwar einer Verletzung der Normalverteilungsvoraussetzung, die Einschränkungen hinsichtlich der Stichprobengröße sind aber die gleichen wie bei der herkömmlichen χ^2-Statistik (Beaujean, 2014). Deshalb sollte ein theoretisch plausibles Modell bei einem signifikanten χ^2-Test nicht direkt zurückgewiesen werden, sondern weitere Indikatoren für die Modellevaluation berücksichtigt werden (Brown, 2015; West, Taylor & Wu, 2012). Trotz dieser Alternativen sollte das Ergebnis des χ^2-Tests, einschließlich Freiheitsgraden, immer berichtet werden (Boomsma, 2000; Kline, 2011).

Alternative Fitstatistiken im Überblick.

Alternative Indikatoren, die entwickelt wurden, um den Einschränkungen des χ^2-Tests zu begegnen, können grob drei Kategorien zugeordnet werden: Absolute Fitindikatoren (*absolute fit indexes*), inkrementelle oder komparative Fitindikatoren (*incremental/comparative fit indexes*) und Modellsparsamkeit (*parsimony*) (Brown, 2015; Weiber & Mühlhaus, 2014). Absolute Fitindikatoren basieren auf einem Vergleich zwischen empirischer und geschätzter Kovarianzmatrix, beurteilen also, wie gut ein Modell die empirische Kovarianzmatrix vorhersagt (Weiber

& Mühlhaus, 2014) und untersuchen die Abweichung von einem perfekten Modell. Ein größerer Wert eines Indikators bedeutet einen schlechteren Modellfit, bei einem Wert von null läge ein perfekt zu den Daten passendes Modell vor (Berning, 2019). Inkrementelle Fitindikatoren messen dagegen die proportionale Verbesserung des Modellfits beim Vergleich eines Modells mit einem restriktiveren geschachtelten Basismodell, i. d. R. ein Modell, in dem alle manifesten Variablen unkorreliert sind (Hu & Bentler, 1999). Indikatoren zur Beurteilung der Sparsamkeit dienen vor allem dem Modellvergleich von zwei oder mehr Modellen. Sie zielen darauf ab, ein Modell auszuwählen, das die beobachteten Daten möglichst effizient darstellt und hierfür die wenigsten Parameter benötigt – somit also ein möglichst sparsames Modell (Beaujean, 2014; Weiber & Mühlhaus, 2014). Nach Möglichkeit sollten zur Modellbeurteilung immer Fitmaße aus allen drei Kategorien herangezogen werden (Weiber & Mühlhaus, 2014). Im Folgenden werden die in dieser Arbeit verwendeten Indikatoren einschließlich der jeweiligen Cutoff-Werte vorgestellt.

Absolute Fitindikatoren.

Ein weit verbreiteter Indikator zur Messung des absoluten Modellfit ist der *Root Mean Square Error of Approximation* (RMSEA, vgl. bspw. Browne & Cudeck, 1992). Der RMSEA überprüft, ob ein spezifiziertes Modell die Daten ausreichend genau wiedergibt. Dies steht im Gegensatz zum χ^2-Test, welcher zur Testung eines exakten Modellfits verwendet wird. Dies bedeutet, die Nullhypothese eines perfekten Modellfits wird durch eine annähernde Passung (*close fit*) ersetzt (Schermelleh-Engel et al., 2003). Der Wertebereich des RMSEA liegt zwischen 0 und 1, kann jedoch auch > 1 werden. Je näher der Wert an 0 liegt, desto besser ist demnach der Fit (Beaujean, 2014). Ein Wert ≤ .05 kennzeichnet einen guten Modellfit, der Wertebereich > .05 bis .08 einen akzeptablen Fit (Finch & French, 2015; Schermelleh-Engel et al., 2003). Üblicherweise wird zusätzlich zum Wert des RMSEA das 90 %-Konfidenzintervall (KI) angegeben, dessen Minimum bestenfalls nahe bei .00 liegt und dessen Maximum .10 nicht überschreiten sollte (Gana & Broc, 2019). Im Falle des Abweichens der analysierten Daten von einer multivariaten Normalverteilung steht im Rahmen der MLR-Schätzung mit lavaan eine robuste Variante des RMSEA nach Brosseau-Liard, Savalei und Li (2012) zur Verfügung. Zur Bewertung dieses robusten RMSEA ($RMSEA_{r)}$ können die etablierten Cutoff-Kriterien herangezogen werden (Savalei, 2018).

Der *Standardized Root Mean Square Residual* (SRMR) stellt einen Residuen basierten Fitindikator dar und fasst die Informationen über die Residuen einer Modellschätzung in einem standardisierten Maß zusammen (Schermelleh-Engel et al., 2003; Urban & Mayerl, 2014). Der

SRMR kann Werte zwischen 0 und 1 annehmen, wobei Werte nahe bei 0 einen besseren Fit postulieren (Beaujean, 2014). Während Hu und Bentler (1999) einen absoluten Cutoff-Wert von .08 vorschlagen, differenzieren Schermelleh-Engel et al. (2003) die Bewertung dieses Kriteriums in Anlehnung an Hu und Bentler (1995). Ein SRMR ≤ .05 bildet demnach einen guten Modellfit ab, ein Wert ≤ .10 einen noch akzeptablen Fit.

Ein weiterer absoluter Fitindikator mit einem gewissen ad hoc Charakter ergibt sich aus dem Quotienten des χ^2-Wertes und den zugehörigen Freiheitsgraden, die χ^2/df-Ratio. Dieses Verhältnis gibt an, wie viel Mal größer der beobachtete χ^2-Wert ggü. dem Erwartungswert von χ^2 mit einer bestimmten Anzahl an Freiheitsgraden ist. Je kleiner der Wert dieses Verhältnisses, desto besser der Fit, die empfohlenen Grenzwerte liegen hierbei bei ≤ 2 bzw. ≤ 3 (Beaujean, 2014; Bollen, 2014). Jedoch besteht bei diesem Kriterium das gleiche Problem einer inflationären Ablehnung des Modells bei zu großen Stichproben wie bei der reinen Betrachtung der χ^2-Statistik (Bollen, 2014; West et al., 2012).

Inkrementelle Fitindikatoren.

Zu den bekanntesten inkrementellen Fitindikatoren zählen der *Comparative Fit Index* (CFI, vgl. Bentler, 1990) und der *Tucker-Lewis-Index* (TLI, vgl. Tucker & Lewis, 1973). CFI und TLI vergleichen das untersuchte Modell mit einem Basis- bzw. Unabhängigkeitsmodell (Kline, 2011). Der CFI kann Werte von 0 bis 1 annehmen, wobei Werte nahe bei 1 für einen guten Modellfit sprechen. Der TLI als nicht normierter Indikator kann theoretisch auch Werte größer 1 oder kleiner 0 annehmen, lässt sich jedoch ebenso wie der CFI interpretieren, wonach Werte nahe bei 1 einen guten Modellfit repräsentieren (Brown, 2015; West et al., 2012). Um von einem akzeptablen Modellfit zu sprechen, sollten CFI und TLI Werte ≥ .950 aufweisen, besser jedoch über .970 liegen (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003). Auch für CFI und TLI bietet lavaan im Rahmen der MLR-Schätzung die robuste Alternative nach Brosseau-Liard und Savalei (2014), im weiteren als CFI_r und TLI_r bezeichnet, wiederum mit denselben Cutoff-Kriterien wie bei den nicht robusten Indikatoren (Savalei, 2018).

Modellsparsamkeit.

Zur Beurteilung der Sparsamkeit bei Modellvergleichen wird in dieser Arbeit das *Akaike's Information Criterion* (AIC, vgl. bspw. Akaike, 1974) herangezogen. Dieser deskriptive Indikator berechnet sich als eine Funktion aus χ^2, korrigiert um die Anzahl der frei zu schätzenden Parameter. Der Wertebereich des AIC ist nicht normiert, die Beurteilung erfolgt ausschließlich im Vergleich zu konkurrierenden Modellen. Der kleinere AIC-Wert kennzeichnet das sparsamere

Modell (Gana & Broc, 2019; Schermelleh-Engel et al., 2003; Tabachnick & Fidell, 2010). Der AIC eignet sich auch zum Vergleich nicht-geschachtelter Modelle (Schreiber, 2008). Alternativ wären das *Consistent AIC* (CAIC, vgl. bspw. Bozdogan, 1987) oder das *Bayesian Information Criterion* (BIC) geeignet, die Modellsparsamkeit zu beurteilen. Der CAIC wird jedoch von lavaan nicht ausgegeben und findet aus diesem Grund keine Berücksichtigung, während die Betrachtung des BIC zu denselben Schlussfolgerungen der Modellbeurteilung führt, wie das Heranziehen des AIC, weshalb auch auf einen zusätzlichen Bericht des BIC verzichtet wird.

Tabelle 16: Cutoff-Werte der verwendeten Fitindikatoren (nach Beaujean, 2014; Bollen, 2014; Hu & Bentler, 1995, 1999; Savalei, 2018; Schermelleh-Engel et al., 2003)

	χ_r^2	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	SRMR	AIC
akzeptabel		≤ 3	$> .950$	$> .950$	$\leq .080$	$\leq .100$	
gut	*n.s.*	≤ 2	$> .970$	$> .970$	$\leq .050$	$\leq .050$	$< AIC_{Alt.}$

Anmerkung. n.s.: nicht signifikant.

Lokale Gütemaße und Parameterschätzungen.

Auf Messmodellebene sollten neben der Beurteilung des Gesamtmodells verschiedene lokale Gütekriterien bestimmt werden. Zur Bestimmung der internen Konsistenz der Faktoren wurde jeweils Cronbachs Alpha (α) (Cronbach, 1951) bestimmt. Zwar liegt der Grenzwert für eine angemessene interne Konsistenz i. d. R. bei $\alpha \geq .70$ (Nunnally, 1978), jedoch können u. U. bei Skalen mit nur zwei oder drei Indikatorvariablen auch Werte ≥ .40 als akzeptabel erachtet werden (Zinnbauer & Eberl, 2004). Als Kriterium zur Überprüfung der Itemtrennschärfe wurde die korrigierte Item-Skala-Korrelation (*Item-to-Total-Korrelation* – KITK) berechnet. Diese gibt an, wie gut ein Item das Gesamtergebnis einer Skala vorhersagt, und sollte einen Wert von ≥ .50 aufweisen (Weiber & Mühlhaus, 2014).

Im Hinblick auf die Parameterschätzungen und Gütebeurteilungen der CFA geben die standardisierten Faktorladungen (λ^s) dabei Hinweise auf die formale Validität eines Faktors und sollten bei mindestens .50 liegen. Jedoch finden sich auch restriktivere Vorgaben, die eine Faktorladung $\geq .70$ verlangen (Urban & Mayerl, 2014). Andere Quellen fordern hingegen von einem akzeptablen Indikator nur ein $\lambda^s \geq .40$ und halten in inhaltlich begründbaren Fällen auch niedrigere Faktorladungen für u. U. tolerierbar (Berning, 2019).

Der Determinationskoeffizient R^2 entspricht der quadrierten standardisierten Faktorladung eines Items und beschreibt den Anteil an durch den zugrundeliegenden Faktor erklärter Varianz eines Items. Grundsätzlich ist es das Ziel, ein Modell zu finden, in dem ein möglichst großer Varianzanteil eines jeden Items durch den zugrundeliegenden Faktor erklärt wird (Bollen, 2014). Ein Richtwert ist dabei optimalerweise eine Varianzaufklärung von mindestens 50 % ($R^2 > .50$) (Kline, 2011).

Eine weitere wichtige Parameterschätzung in Strukturgleichungsmodellen ist der Standardfehler (S. E. – *Standard Error*), welcher die Variabilität der Parameterschätzungen misst. Ein Standardfehler steht für die Varianz der einzelnen Schätzwerte und dient dadurch als Indikator für die Stabilität der Schätzungen. Da es für die Bewertung von Standardfehlern kein eindeutiges normiertes Kriterium gibt, können diese lediglich nach Auffälligkeiten inspiziert werden. Hierbei gilt es auffällig große oder kleine Werte – insbesondere im Verhältnis zu anderen Indikatoren – zu identifizieren, weil diese Hinweise auf diverse Modellmissspezifikationen geben können (Brown, 2015; Gana & Broc, 2019; Urban & Mayerl, 2014). Kleine Standardfehler indizieren zwar eine akkurate Parameterschätzung, jedoch kann bei einem Standardfehler von null keine Signifikanz einer Parameterschätzung berechnet werden. Zu große Standardfehler wiederum können u. a. zu Problemen bei der Schätzgenauigkeit führen (Brown, 2015) und bspw. in nicht-signifikanten, aber hohen Koeffizientenschätzungen resultieren (Urban & Mayerl, 2014).

Die durchschnittlich erfasste Varianz (DEV) gibt an, wieviel Varianz der Indikatoren durchschnittlich durch das latente Konstrukt erklärt wird. Der Schwellenwert dieses Indikators für eine gute Reliabilität liegt bei DEV $\geq .50$ (Bagozzi, 1991; Fornell & Larcker, 1981; Weiber & Mühlhaus, 2014). Weiterhin kann die DEV zur Bestimmung der Diskriminanzvalidität herangezogen werden. Zusätzlich sollte die Korrelation zwischen zwei latenten Variablen bestenfalls nicht größer als .85 (Brown, 2015) bzw. < .90 (Kline, 2011) sein. Diskriminanzvalidität (nach Campbell & Fiske, 1959) liegt vor, wenn sich die Messungen verschiedener Konstrukte unterscheiden und davon ausgegangen werden kann, dass zwei latente Faktoren tatsächlich zwei getrennte Konstrukte repräsentieren (Hair, Black, Babin & Anderson, 2010).

Schritt 6: Respezifikation/Modifikation der Modellstruktur

Ein Modell, das einen unzureichenden Modellfit oder sonstige Unzulänglichkeiten aufweist, kann basierend auf statistischen und inhaltlich rationalen Überlegungen überarbeitet werden

(Kline, 2011). Dabei gilt zu beachten, dass die Methode der Strukturgleichungsmodellierung dadurch ihren konfirmatorischen Charakter verliert, da eine nachträgliche Modellanpassung eher einem explorativen Vorgehen entspricht. Daher sollte ein nachträglich angepasstes Modell bestenfalls mit einer neuen unabhängigen Stichprobe validiert werden (Ullman & Bentler, 2003). In dieser Arbeit erfolgte entsprechend diesem Vorgehen in Kapitel 5.2 eine Anpassung des ursprünglich aufgestellten Modells, welches dann in Kapitel 5.3 mit Hilfe einer zweiten unabhängigen Stichprobe validiert wurde.

Schritt 7: Replikation

Kline (2011) mahnt an, dass Strukturgleichungsmodelle viel zu selten mit neuen Daten repliziert werden. Eine solche Replikationsstudie wurde jedoch in dieser Arbeit durchgeführt, indem das untersuchte (angepasste) Modell mit einer Stichprobe aus Sekundarschullehrkräften im Kontext von VERA8 mit Hilfe einer latenten Gruppenanalyse reproduziert wurde (siehe Kapitel 5.4). Das Konzept und die Vorgehensweise latenter Gruppenanalysen werden im folgenden Kapitel beschrieben.

4.3.4. Latente Gruppenanalyse

Konzept von Gruppenanalysen

Zur Untersuchung der Gültigkeit des untersuchten Modells in einer weiteren Stichprobe aus VERA8-Lehrkräften und zur Ermittlung möglicher Gruppenunterschiede wurde im nächsten Schritt ein Mehrgruppenmodell analysiert. Die Besonderheit einer solchen Gruppenanalyse liegt hier in der simultanen Schätzung des Modells für alle analysierten Gruppen (Gana & Broc, 2019). Gruppenanalysen bzw. Gruppenvergleiche eignen sich für die Beantwortung der folgenden Fragen: (1) Messen die gewählten Indikatorvariablen in verschiedenen Gruppen (z. B. Geschlecht, Altersgruppen) dasselbe Konstrukt, d. h. ist ein definiertes Messmodell über bestimmte Gruppen hinweg gültig? Ist dies gegeben, liegen eine Messinvarianz bzw. Messäquivalenz vor. Das Vorliegen der Messinvarianz der latenten Konstrukte bildet die Voraussetzung für die Untersuchung folgender weiterer Fragestellungen: (2) Besitzen die definierten Strukturbeziehungen in unterschiedlichen Gruppen Gültigkeit und inwiefern gleichen oder unterschieden sich die Wirkungsstärken der Effekte? (3) Gibt es (signifikante) Unterschiede in den Mittelwerten der latenten Variablen zwischen den analysierten Gruppen (Weiber & Mühlhaus,

2014)? In dieser Arbeit war diesbezüglich die Übertragbarkeit der genutzten Messmodelle auf eine Stichprobe von Lehrkräften der Sekundarstufe (VERA8) von Interesse, ebenso wie die Analyse potenzieller Unterschiede der latenten Mittelwerte und Strukturbeziehungen zwischen den Grund- und Sekundarschullehrkräften.

Da im Rahmen dieser Untersuchung unterschiedlich große Gruppen mit Hilfe eines Gruppenmodells analysiert werden (siehe Kapitel 5.4), wird an dieser Stelle darauf hingewiesen, dass die Übereinstimmung bzw. Annäherung der Gruppengrößen keine Voraussetzung für die Schätzung eines Gruppenmodells darstellt (Schwab & Helm, 2015). In einer Simulationsstudie zeigt sich kein Einfluss unterschiedlicher Gruppengrößen auf das Ergebnis einer Invarianzprüfung, nicht einmal bei einem Verhältnis zweier Gruppen von 1:4 (Koh & Zumbo, 2008), während andere Arbeiten sogar bei einem Gruppenverhältnis von ca. 1:9 eine Invarianzanalyse durchführen (Schwab & Helm, 2015).

Messinvarianzprüfung

Im Zuge der Beantwortung der ersten Fragestellung gruppenanalytischer Untersuchungen ist es das Ziel einer Messinvarianzprüfung zunächst herauszufinden, ob ermittelte Gruppenunterschiede einer latenten Variablen tatsächlich auf Unterschiede zwischen den betrachteten Gruppen oder etwa auf das verwendete Messinstrument zurückzuführen sind (Beaujean, 2014) sowie im Sinne einer Konstruktvalidierung die Unabhängigkeit der Struktur und Metrik der analysierten Konstrukte von der jeweiligen Stichprobe darzulegen (Gana & Broc, 2019). Die Frage nach dem Vorliegen von Messinvarianz lässt sich jedoch nicht in eine binäre Kategorie einordnen, sondern ist auf einem Kontinuum zu verorten (Bollen, 2014). Hierbei spricht man von hierarchisch geschachtelten Modellen (*nested models*), bei denen ein Modell mit bestimmten restringierten Parametern in ein Modell, in welchem diese Parameter frei geschätzt werden, geschachtelt ist. Zwei Modelle gelten dann als geschachtelt, wenn die freien Parameter des restriktiveren Modells eine Teilmenge der freien Parameter des weniger restriktiven Modells darstellen (Schermelleh-Engel et al., 2003). Zur empirischen Überprüfung der Messinvarianz werden dementsprechend zunächst die Modelle in den einzelnen Gruppen separat getestet, bevor ein gruppenübergreifendes Modell analysiert wird und nacheinander die verschiedenen Invarianzbedingungen getestet werden, indem zunehmend restriktivere Modelle spezifiziert und geschätzt werden (Brown, 2015; Millsap & Olivera-Aguilar, 2012; Ullman & Bentler, 2003). Die Invarianzprüfung wird jeweils nur fortgesetzt, wenn die vorherige weniger restriktive Invarianzbedingung erfüllt ist. Ist dies nicht der Fall, ist eine weitere Prüfung nicht sinnvoll (Bollen,

2014; Brown, 2015). Ziel ist es hierbei durch zunehmende Restriktionen den Modellfit nicht wesentlich zu verschlechtern (Ullman & Bentler, 2003) (siehe hierzu auch Beurteilungskriterien der Invarianz, S. 159f.).

Stufen der Messinvarianz.

In der Methodenliteratur werden die im Folgenden aufgeführten Stufen der Invarianz unterschieden, wobei die verwendete Terminologie häufig uneinheitlich ist. Die Testung der Invarianz erfolgt typischerweise entsprechend dieser Reihenfolge (vgl. bspw. Beaujean, 2014; Brown, 2015; Cheung & Rensvold, 2002; Finch & French, 2015; Gana & Broc, 2019; Keith, 2019; Meredith, 1993; Millsap & Olivera-Aguilar, 2012; Reinecke, 2014):

1. Konfigurale Invarianz: Diese schwächste Form faktorieller Invarianz bezieht sich auf die Gültigkeit einer gemeinsamen Faktorstruktur in allen analysierten Gruppen. Daher wird in dem entsprechenden ersten Prüfschritt zunächst untersucht, ob das untersuchte Modell in allen Gruppen zu den Daten passt und entsprechend die Indikatorvariablen in allen Gruppen die gleichen Konstrukte messen. Annahmen zur Gleichheit der Parameterschätzungen werden jedoch nicht getroffen. Die Gültigkeit des geschätzten Gruppenmodells wird anhand der üblichen Fitstatistiken beurteilt. Diese basale Form der Invarianz ist Voraussetzung für die Überprüfung aller weiteren Formen der Invarianz. Erweist sich eine angenommene Faktorstruktur als nicht passend für alle betrachteten Gruppen, ist eine weitere Untersuchung gruppenbezogener Unterschiede nicht sinnvoll.

2. Metrische Invarianz: Diese restriktivere Form der Invarianz, auch schwache faktorielle Invarianz oder Messinvarianz genannt, unterstellt, ebenso wie die konfigurale Invarianz, die Gültigkeit einer gemeinsamen Faktorstruktur in verschiedenen Gruppen und verlangt zusätzlich die Gleichheit der Faktorladungen zwischen den Gruppen. Zur Überprüfung metrischer Invarianz werden daher die standardisierten Faktorladungen zwischen den Gruppen gleichgesetzt, während alle anderen Parameter frei geschätzt werden. Sodann wird der Modellfit dieses restriktiveren Modells bzw. die Veränderung der Modellpassung gegenüber dem vorherigen weniger restriktiven Modell beurteilt (siehe Tabelle 17, S. 160). Ist diese Invarianzbedingung erfüllt, kann von einer inhaltlich gleichen Bedeutung der latenten Konstrukte in den einzelnen Gruppen ausgegangen werden. Diese Stufe der Invarianz bildet die minimale Voraussetzung zur gruppenübergreifenden vergleichenden Analyse der Strukturbeziehungen zwischen den latenten Konstrukten in einem SEM.

3. Skalare Invarianz: Das Vorliegen skalarer Invarianz, auch starke faktorielle Invarianz genannt, setzt neben den Anforderungen metrischer Invarianz zusätzlich die Gleichheit der Indikatorkonstanten (Intercepts) voraus. Diese Stufe der Invarianz gilt als Voraussetzung für eine Analyse latenter Mittelwertunterschiede. Ein unter der Voraussetzung skalarer Invarianz ermittelter latenter Mittelwertunterschied bedeutet einen tatsächlichen Unterschied auf Konstruktebene, keinen bspw. durch die Messung in einer Gruppe verursachten Unterschied. Skalare Invarianz impliziert, dass Mittelwertdifferenzen der Indikatorvariablen auf Mittelwertunterschiede des gemeinsamen Faktors, genauer des latenten Konstrukts, zurückzuführen sind. Die Testung skalarer Invarianz erfolgt durch die, im Vergleich zum Modell metrischer Invarianz zusätzliche Gleichsetzung der Intercepts der Indikatorvariablen. Zur Beurteilung wird der Modellfit mit dem des metrischen Modells abgeglichen (siehe Tabelle 17, S. 160).

4. Strikte (Faktor-)Invarianz: Der letzte Aspekt der Messinvarianz setzt zusätzlich die Gleichheit der Messfehler zwischen Gruppen voraus, weshalb auch häufig von Residual- oder Messfehler-Invarianz gesprochen wird. Für die Prüfung dieser Invarianzstufe werden daher zusätzlich zu den bisherigen Restriktionen die Residuen der Indikatorvariablen zwischen den untersuchten Gruppen fixiert und der Modellfit im Vergleich zu dem Modell metrischer Invarianz beurteilt. Im Gegensatz zu metrischer bzw. skalarer Invarianz ist das Vorliegen einer strikten Invarianz jedoch keine notwendige Voraussetzung für bestimmte Gruppenvergleiche wie Mittelwertanalysen und die Untersuchung von Strukturparametern und besitzt daher nur eingeschränkt praktische Relevanz.

Darüber hinaus gibt es noch weitere striktere Formen der Invarianz, wie die Gleichheit der Varianzen und Kovarianzen, deren Überprüfung jedoch einen nur sehr begrenzten Erkenntnisgewinn liefert (Beaujean, 2014; Kline, 2011) und daher auch in dieser Arbeit nicht weiter behandelt wird.

Partielle Messinvarianz.

Da strenge Invarianzbedingungen im Wesentlichen idealtypische theoretische Annahmen abbilden, deren Erfüllung sich in der Praxis häufig als unrealistisch erweist, wird in der praktischen Anwendung häufig auf das Konzept der partiellen Invarianz zurückgegriffen (vgl. Byrne, Shavelson & Muthén, 1989). Wird ein bestimmtes Level der Invarianz nicht erreicht, muss ein Modell nicht zwangsläufig verworfen und die Prüfung weiterer Invarianzstufen beendet werden. Stattdessen können mit Hilfe von Modifikationsindizes und inhaltlichen bzw.

theoriegeleiteten Überlegungen einzelne nicht invariante Parameter identifiziert und das Modell entsprechend durch die (schrittweise) Aufhebung einzelner Restriktionen angepasst und erneut getestet werden (Brown, 2015; Gana & Broc, 2019; Keith, 2019; Millsap & Olivera-Aguilar, 2012).

Sind bspw. die Voraussetzungen metrischer Invarianz nicht erfüllt, können die Faktorladungen einzelner (oder auch mehrerer) Indikatorvariablen freigesetzt werden, während die übrigen Ladungen zwischen den Gruppen gleichgesetzt bleiben. Bei den anderen Invarianzstufen kann analog verfahren werden (Keith, 2019; Kline, 2011). In einem Modell mit partieller skalarer Invarianz werden demnach einzelne Intercepts frei geschätzt, während die übrigen entsprechend gleichgesetzt sind (Millsap & Olivera-Aguilar, 2012). Auch unter der Voraussetzung partieller Invarianz ist somit eine Testung weiterer Invarianzformen möglich, ebenso wie speziell die Analyse von Mittelwertunterschieden bei nur partieller skalarer Invarianz (Byrne et al., 1989; Kline, 2011).

Zur Frage, wie viele Indikatoren eines Konstrukts nicht invariante Parameterschätzungen aufweisen sollten, um als hinreichend im Sinne einer angemessenen Repräsentation vollständiger Invarianz zu gelten, gibt es in der Literatur nur wenig konkrete Aussagen, weshalb diese schwierig zu beantworten ist (Millsap & Olivera-Aguilar, 2012). Konservative Einschätzungen empfehlen jeweils nur eine invariante Faktorladung oder Intercept zuzulassen (Keith, 2019). Andere Autor*innen fordern weniger rigide jeden Faktor mit mindestens zwei metrisch bzw. skalar invarianten Indikatorvariablen zu messen (Nusser, Carstensen & Artelt, 2015) oder dass der Anteil invarianter Indikatoren überwiegen sollte (Putnick & Bornstein, 2016). Urban und Mayerl (2014) merken an, dass nur die besonders wichtigen Items eines Konstrukts zwingend invariant sein sollten. Generell sollte die Modellierung partieller Invarianz als post hoc Verfahren immer mit Bedacht angewendet und auch bei der Ergebnisinterpretation berücksichtigt werden (Brown, 2015; Vandenberg & Lance, 2000).

Beurteilungskriterien der Invarianz.

Der klassische Ansatz zur Evaluierung geschachtelter Modelle allgemein und im speziellen zur Bewertung der Messinvarianz zwischen verschiedenen Gruppen ist der statistische Ansatz mit Hilfe eines χ^2-Differenztests. Dieser basiert auf der Veränderung der χ^2-Statistik beim Vergleich zweier Modelle mit zunehmend invarianten Parameterschätzungen. Verändert sich durch weitere Modellrestriktionen der χ^2-Wert ($\Delta\chi^2$) nicht signifikant, kann von dem Vorliegen der getesteten Invarianzstufe ausgegangen werden und die Invarianzhypothese muss nicht

verworfen werden (Beaujean, 2014; Schermelleh-Engel et al., 2003). Für die Beurteilung geschachtelter Modelle und die Invarianzanalyse mit Hilfe eines χ^2-Differenztests gelten jedoch die gleichen Stichprobengrößen abhängigen Restriktionen wie für den χ^2-Test (Cheung & Rensvold, 2002; Kline, 2011; Meade, Johnson & Braddy, 2008).

Als Alternative zum χ^2-Differenztest hat sich daher die Verwendung des weniger Stichprobengrößen sensitiven CFI (Comparative Fit Index) etabliert. Analog zum χ^2-Differenztest wird dessen Veränderung (ΔCFI) zwischen zwei hierarchisch geschachtelten Modellen betrachtet. ΔCFI sollte dabei i. d. R. .010 nicht überschreiten, bei einer Verschlechterung des Modellfits um $\Delta CFI \leq -.010$ wird die Invarianzhypothese nicht zurückgewiesen (Cheung & Rensvold, 2002; Meade et al., 2008; Vandenberg & Lance, 2000). Bei einer weniger konservativen Betrachtung kann auch ein Cutoff-Wert von $\Delta CFI \leq -.020$ als noch akzeptabel erachtet werden, jedoch sollte bei einer Differenz zwischen -.010 und -.020 das Vorliegen tatsächlicher Gruppenunterschiede in Betracht gezogen werden (Vandenberg & Lance, 2000).

Als weitere Kriterien zur Invarianzbeurteilung eignen sich die Veränderungen des RMSEA und SRMR (ΔRMSEA bzw. ΔSRMR). Weil diese beiden Indikatoren aber einige Limitationen, wie die Abhängigkeit des RMSEA von der Modellkomplexität, aufweisen, empfiehlt Chen (2007) die Verwendung von ΔCFI als Hauptkriterium zur Prüfung der Invarianz und die Betrachtung von ΔRMSEA und ΔSRMR als ergänzende Informationen. Die entsprechenden Cutoff-Werte sind in Tabelle 17 dargestellt. In dieser Arbeit berechnet sich ΔCFI, wie bereits in Kapitel 4.3.3 beschrieben, unter Verwendung des robusten CFI und wird daher im Weiteren als ΔCFI_r bezeichnet, ΔRMSEA entsprechend als $\Delta RMSEA_r$.

Tabelle 17: Cutoff-Kriterien zur Überprüfung faktorieller Invarianz (nach Cheung & Rensvold, 2002; Meade et al., 2008; Vandenberg & Lance, 2000)

	Gesamtmodellfit	ΔCFI_r	$\Delta RMSEA_r$	ΔSRMR
Konfigurale Invarianz	akzeptabel			
Metrische Invarianz	akzeptabel	≤ -.010	≤ .015	≤ .030
Skalare Invarianz	akzeptabel	≤ -.010	≤ .015	≤ .010
Strikte Invarianz	akzeptabel	≤ -.010	≤ .015	≤ .010

Neben der Überprüfung des Ausmaßes einer möglichen Modellverschlechterung durch restriktivere Parameterschätzungen muss zusätzlich immer die Gesamtgüte des Modells beurteilt

werden, da es durchaus möglich ist, dass zwar Messinvarianz vorliegt, ein Modell dennoch insgesamt einen nur unzureichenden Modellfit aufweist. Im Falle eines nicht fittenden Modells wäre eine Invarianzprüfung daher nicht zielführend (Vandenberg & Lance, 2000). Ein zumindest akzeptabler Modellfit gemäß der in Tabelle 16 (S. 153) aufgeführten Kriterien, ist somit immer Voraussetzung aller Stufen faktorieller Invarianz.

Gruppenanalysen auf Strukturebene

Sind bestimmte Bedingungen der Messinvarianz erfüllt, kann die Invarianzprüfung auf Ebene des Strukturmodells fortgesetzt werden. Hierbei können Strukturkoeffizienten, latente Mittelwerte und deren Varianzen bzw. Residualvarianzen auf Invarianz untersucht werden (Reinecke, 2014; Vandenberg & Lance, 2000). Hierbei geht es i. d. R. um die Beantwortung spezifischer Forschungsfragen, nicht nur wie bei der Invarianzprüfung um die Frage, ob ein Messinstrument in verschiedenen Gruppen funktioniert (Keith, 2019). Ein Ziel der vorliegenden Arbeit ist es, mögliche abweichende Einschätzungen von Lehrkräften in Grundschulen und weiterführenden Schulen zu Vergleichsarbeiten zu untersuchen und etwaige Unterschiede in den postulierten Kausalbeziehungen aufzudecken. Daher sind für diese Arbeit, entsprechend der zu Beginn des Abschnitts aufgeführten Fragestellungen 2 und 3, die Analyse von Strukturbeziehungen und latenten Mittelwerten von Bedeutung und werden im Folgenden skizziert.

Latente Gruppenmittelwerte können beim Vorliegen zumindest partieller skalarer Invarianz untersucht werden (Beaujean, 2014). Ein Vorteil einer mit einer latenten Mehrgruppenanalyse einhergehenden Untersuchung von Mittelwertdifferenzen zwischen zwei oder mehreren Gruppen gegenüber herkömmlichen Verfahren wie einer ANOVA oder einem t-Test besteht u. a. darin, dass zunächst durch eine Invarianzanalyse überprüft wird, ob Gruppenvergleiche überhaupt angemessen sind. Traditionelle Verfahren unterstellen dagegen einfach eine perfekte Reliabilität, testen aber nicht, ob die Messung tatsächlich in verschiedenen Gruppen auch das Gleiche misst. Durch die latente Modellierung können Messfehler, korrelierte Residuen etc. berücksichtigt werden (Brown, 2015; Keith, 2019; Müller & Schäfer, 2017; Vandenberg & Lance, 2000).

Die Ausprägung von Mittelwertunterschieden in einem latenten Modell wird ermittelt, indem die Intercepts, also die Mittelwerte der latenten Faktoren in einer Gruppe (Referenzgruppe), auf null fixiert und in den anderen Gruppen frei geschätzt werden. Da für diese Testung das Vorliegen (partieller) skalarer Invarianz vorausgesetzt wird, werden zusätzlich Faktorladungen und

Intercepts der Indikatorvariablen zwischen den Gruppen gleichgesetzt. Die frei geschätzten Werte auf latenter Ebene in den anderen Gruppen repräsentieren die Abweichung des latenten Mittelwerts von dem der Referenzgruppe (Brown, 2015; Millsap & Olivera-Aguilar, 2012; Reinecke, 2014). Die Differenz der latenten Mittelwerte wird für die Vergleichsgruppen in lavaan jeweils als Wert des Intercepts ausgegeben (Beaujean, 2014; Gana & Broc, 2019). Latente Mittelwerte werden demnach nicht absolut ermittelt, sondern spiegeln die Unterschiede zwischen Gruppen auf Konstruktebene wider (Reinecke, 2014).

Unterschiede in den Strukturbeziehungen zwischen Gruppen können unter der Voraussetzung mindestens metrischer Invarianz analysiert werden. Hierfür werden die spezifizierten Pfadkoeffizienten eines Strukturmodells zwischen den betrachteten Gruppen gleichgesetzt (Kline, 2011). Zur Bewertung dieser strukturellen Invarianzstufe schlägt Keith (2019), wie bei der Prüfung der Messinvarianz, die Differenz des CFI mit $\Delta CFI \leq .010$ vor.

5. Empirischer Teil

5.1. Überprüfung des theoretischen Modells (Modell 1)

Das folgende Kapitel beschäftigt sich mit der Prüfung des in Kapitel 3, Abbildung 7 aufgestellten Forschungsmodells. In Vorbereitung der Modellschätzung werden zunächst in Kapitel 5.1.1 die deskriptiven Statistiken der Indikatorvariablen der untersuchten Konstrukte berichtet und auf mögliche Auffälligkeiten inspiziert. Im ersten Schritt der Modellprüfung wurden dann die Messmodelle mit Hilfe konfirmatorischer Faktorenanalysen anhand der Daten der VERA3-Evaluation 2018 geschätzt. Neben der Gesamtbeurteilung der Messmodelle wurden zunächst für eine erste Inspektion der untersuchten Skalen die Interkorrelationen der Indikatorvariablen sowie ausgewählte Reliabilitätsindikatoren der ersten Generation (vgl. Weiber & Mühlhaus, 2014) betrachtet. Die Ergebnisse dieser Analysen sind in Kapitel 5.1.2 dokumentiert. In Kapitel 5.1.3 folgt die Analyse der Strukturbeziehungen hinsichtlich der aufgestellten Hypothesen.

5.1.1. Deskriptive Statistiken

In Tabelle 18 sind die deskriptiven Statistiken der manifesten Indikatorvariablen der untersuchten Konstrukte dargestellt. Die Tabelle enthält neben den Itemmittelwerten und Standardabweichungen den prozentualen Anteil fehlender Werte. Zur Beurteilung der multivariaten Normalverteilung der Indikatorvariablen, welche eine Voraussetzung der ML-Schätzmethode wäre (siehe bspw. Beaujean, 2014), wurden zusätzlich Schiefe und Kurtosis sowie das jeweilige Ergebnis des Shapiro-Wilk-Tests (S-W-Test) angegeben. Der Shapiro-Wilk-Test (nach Shapiro & Wilk, 1965; Shapiro, Wilk & Chen, 1968) testet die Nullhypothese, dass eine Normalverteilung vorliegt, gegen die Alternativhypothese, dass keine Normalverteilung besteht. Wird der S-W-Test signifikant, muss die Hypothese der Normalverteilung verworfen werden.

Es zeigt sich ein insgesamt negativer Trend in der Beantwortung der einzelnen Items. Die Mittelwerte der Items des Konstrukts Nutzungsintention liegen bis auf Item NI5 alle unter dem theoretischen Mittelwert von 2.50. Ein ähnliches Bild zeigt sich bei den Items des Konstrukts Einstellung, die mit $M_{AE5} = 2.31$ ($SD = 0.95$) bis $M_{AE3} = 2.41$ ($SD = 0.89$) alle das theoretische Mittel unterschreiten. Die Bewertung der Aufwand-Nutzen Items fällt insgesamt etwas positiver aus, hier liegen zumindest drei Items oberhalb von 2.50. Das inverse Item AN1, das für

diese und alle folgenden Analysen umgepolt wurde (= AN1r), liegt mit 2.06 (*SD* = 0.88) dagegen deutlich unter dem theoretischen Mittel und weist somit eine stark negative Bewertungstendenz auf.

Tabelle 18: Deskriptive Statistiken der Indikatorvariablen der latenten Konstrukte (VERA3 2018)

	Item	*M*	*SD*	Schiefe	Kurtosis	S-W-Test	Anteil fehlender Werte
Nutzungsintention Min = 1, Max = 4 M_T = 2.50	NI1	2.36	0.82	-0.12	-0.68	.86	3.0 %
	NI2	2.38	0.81	-0.14	-0.62	.86	3.3 %
	NI3	2.14	0.77	0.22	-0.42	.85	2.9 %
	NI4	2.37	0.85	0.01	-0.68	.87	2.4 %
	NI5	2.64	0.86	-0.41	-0.46	.85	2.0 %
Einstellung Min = 1, Max = 4 M_T = 2.50	AE1	2.38	0.90	0.01	-0.83	.88	1.0 %
	AE2	2.40	0.86	-0.04	-0.70	.87	1.4 %
	AE3	2.41	0.89	-0.05	-0.79	.87	1.4 %
	AE4	2.38	0.90	-0.03	-0.82	.87	0.8 %
	AE5	2.31	0.95	0.10	-0.97	.87	1.9 %
Aufwand-Nutzen Min = 1, Max = 4 M_T = 2.50	AN1r [a]	2.06	0.88	-0.20	-1.09	.84	3.7 %
	AN2	2.41	0.84	-0.03	-0.62	.87	3.7 %
	AN3	2.35	0.84	0.04	-0.64	.87	3.4 %
	AN4	2.82	0.99	-0.40	-0.89	.86	3.8 %
	AN5	2.80	0.99	-0.38	-0.88	.86	5.6 %
	AN6	2.53	1.02	-0.04	-1.12	.88	3.7 %
Zeitliche Belastung Min = 1, Max = 5 (inv. Polung) M_T = 3.00	ZB1	2.75	0.95	-0.04	-0.13	.89	1.1 %
	ZB2	3.02	0.73	0.01	1.72	.79	0.8 %
	ZB3	3.42	0.83	0.26	0.07	.84	0.6 %
Wahrgenommene Nützlichkeit Min = 1, Max = 4 M_T = 2.50	WN1	2.44	0.84	-0.11	-0.63	.87	4.9 %
	WN2	2.39	0.89	-0.01	-0.78	.87	4.4 %
	WN3	2.48	0.90	-0.09	-0.79	.88	3.9 %
	WN4	2.54	0.90	-0.17	-0.76	.87	4.7 %
	WN5	2.64	0.91	-0.29	-0.68	.87	3.5 %

Anmerkungen. N = 4 141 (VERA3 2018); M_T: Theoretischer Mittelwert; inv. Polung: inverse Polung; [a] Das negativ gepolte Item AN1 wurde für die Berechnungen umgepolt (= AN1r); *p*-Werte der Shapiro-Wilk-Tests (S-W-Test): $p < .001$.

Bei dem Konstrukt zeitliche Belastung zeigt sich ein im Vergleich zu den anderen Konstrukten differenzierteres Bild: Während Item ZB1 mit einem Mittelwert von 2.75 (*SD* = 0.95) unterhalb des theoretischen Mittelwerts von 3.00 dieser 5-stufigen Skala liegt und somit aufgrund der inversen Polung eine insgesamt eher niedrige zeitliche Belastung anzeigt und Item ZB2 mit *M* = 3.02 (*SD* = 0.73) relativ neutral bewertet wird, zeigt Item ZB3 mit 3.42 (*SD* = 0.83) eine eher große zeitliche Belastung an. Die Mittelwerte der Items der wahrgenommenen Nützlichkeit liegen zwischen M_{WN2} = 2.39 (*SD* = 0.89) und M_{WN5} = 2.64 (*SD* = 0.91). Die ersten drei Items weisen dabei eine insgesamt eher negative Bewertungstendenz auf, die Items WN4 und WN5 eine leicht positive.

Der Anteil fehlender Werte der einzelnen Items liegt zwischen 0.6 % (Item ZB3) und maximal 5.6 % (Item AN5). Auffällig ist hier, dass die Items derjenigen Konstrukte, deren Bewertung unmittelbar auf Basis der zum Zeitpunkt der Evaluationsbefragung verfügbaren Informationen möglich war, nämlich Einstellung und zeitliche Belastung, durchschnittlich weniger fehlende Werte aufweisen als die Items der übrigen Konstrukte. Dies könnte ein Indikator dafür sein, dass einige Lehrkräfte Schwierigkeiten hatten, die Items dieser Konstrukte vor dem Erhalt der Rückmeldungen zu beantworten.

Die Verteilungsmaße Schiefe und Kurtosis weisen bei nahezu allen Items auf eine Abweichung der Werte der Indikatorvariablen von der Normalverteilung hin. Zusätzlich zeigt die Überprüfung der Normalverteilung mittels des Shapiro-Wilk-Tests, dass bei allen Items eine Abweichung von der Normalverteilung vorliegt und die Nullhypothese der Normalverteilung verworfen werden muss. Die Ergebnisse der Verteilungsbeurteilung sprechen somit gegen das Vorliegen einer multivariaten Normalverteilung und daher, wie in Kapitel 4.3.3 beschrieben, für die Verwendung der MLR-Schätzmethode für die folgende latente Strukturgleichungsanalyse und alle CFAs. Die fehlenden Werte, unter Annahme eines MAR, wurden bei allen Modellschätzungen durch die FIML-Methode berücksichtigt.

5.1.2. Skalenanalysen – Konfirmatorische Faktorenanalyse (VERA3 2018)

Nutzungsintention

Für das Konstrukt Nutzungsintention sind in Tabelle 19 die manifesten Korrelationen der Indikatorvariablen, korrigierte Item-Skala-Korrelation, Cronbachs Alpha sowie die

durchschnittlich erfasste Varianz dargestellt. Alle Interkorrelationen der Indikatorvariablen der Skala Nutzungsintention erweisen sich bei einer zweiseitigen Testung als signifikant ($p < .001$). Die Höhe der Korrelationskoeffizienten liegt zwischen .52 und .75 und ist somit als hoch einzuschätzen (Cohen, 1988). Die Werte der korrigierten Item-Skala-Korrelation liegen für alle Items über dem kritischen Wert von .50, Cronbachs Alpha bei .89. Die durchschnittlich erfasste Varianz liegt bei .61 und somit über dem Grenzwert von .50. Insgesamt weisen die Kennwerte auf eine gute interne Konsistenz des Konstrukts hin.

Tabelle 19: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Nutzungsintention (VERA3 2018)

		1.	2.	3.	4.	5.	α	DEV
1.	NI1	**.86**						
2.	NI2	.75	**.84**					
3.	NI3	.73	.70	**.83**			.89	.61
4.	NI4	.61	.57	.60	**.71**			
5.	NI5	.55	.57	.54	.52	**.67**		

Anmerkungen. $N = 4\,141$ (VERA3 2018); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen; alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test).

Im Anschluss an diese erste Inspektion der Items wurde das Messmodell des Konstrukts Nützlichkeit (siehe Abbildung 12) mit einer konfirmatorischen Faktorenanalyse geschätzt. Zur Skalierung der Metrik der konfirmatorischen Faktorenanlyse wurde die erste unstandardisierte Faktorladung auf 1 fixiert, wodurch die latente Variable die Metrik des ersten Indikators übernimmt. Dieses Vorgehen ist in den meisten Statistikprogrammen, so auch in lavaan, als default Einstellung implementiert (Berning, 2019; Bollen, 2014) und wurde in dieser Arbeit für alle Messmodelle beibehalten. Dies führt dazu, dass die fixierten Indikatoren einen Standardfehler von .000 haben und nicht auf statistische Signifikanz getestet werden können (Brown, 2015; Kline, 2011).

Die Fitstatistiken des Modells sind in Tabelle 20 dargestellt. Nahezu alle Fitindikatoren weisen auf eine gute bzw. akzeptable ($RMSEA_r$) Modellpassung hin. Lediglich die robuste Schätzung des Messmodells ergibt einen χ_r^2-Wert von 83.92 ($df = 5$) und erweist sich als signifikant ($p < .001$), weshalb die Hypothese eines exakten Modellfits anhand der χ_r^2-Statistik zurückgewiesen werden muss. Aufgrund der Stichprobengröße von $N = 4141$ kann dieser

Fitindikator jedoch, wie in Kapitel 4.3.3 (Schritt 5: Modellevaluation) ausgeführt, vernachlässigt werden (Brown, 2015; Schermelleh-Engel et al., 2003; West et al., 2012). Gleiches gilt für die χ_r^2/df-Ratio, die für dieses Modell deutlich über dem Grenzwert von 3 liegt. Da die Problematik der Stichprobengröße und damit einhergehend der Signifikanz von χ_r^2 bei allen Modellschätzungen in dieser Arbeit besteht, wird im Folgenden darauf verzichtet, χ_r^2 und die χ_r^2/df-Ratio näher zu beschreiben. Die Werte werden lediglich der Vollständigkeit halber in den Tabellen mit den Fitstatistiken der jeweiligen Modelle aufgeführt, aber nicht weiter erläutert.

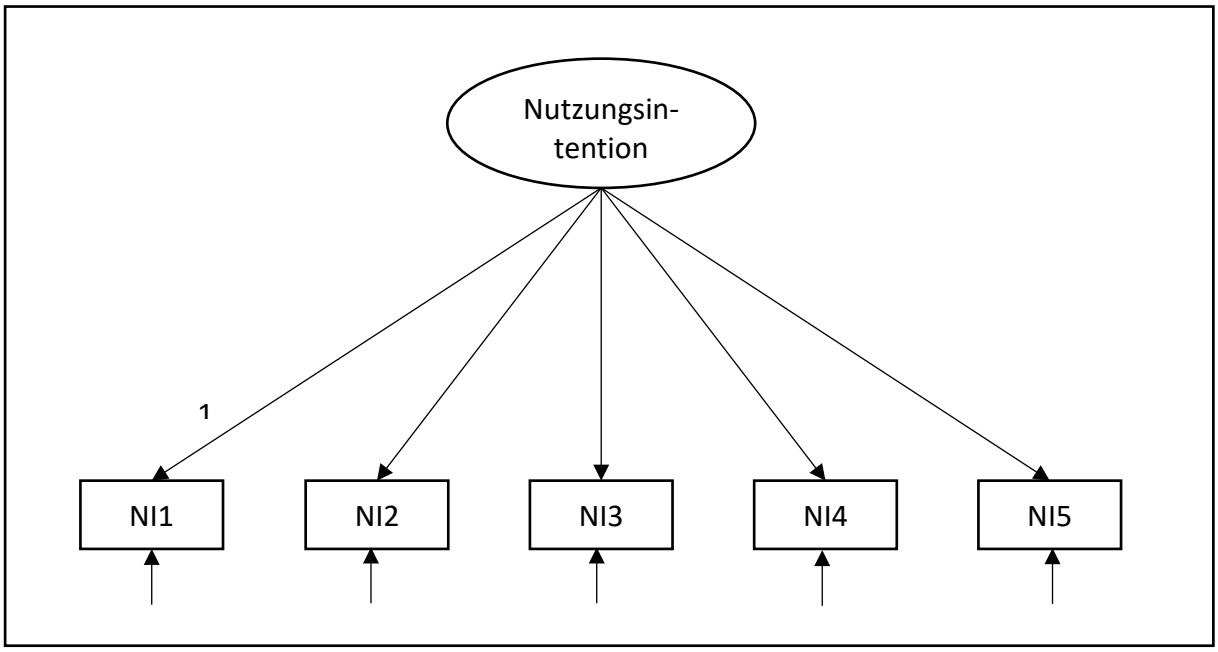


Abbildung 12: Messmodell des Konstrukts Nutzungsintention

Tabelle 20: Fitstatistiken der konfirmatorischen Faktorenanalyse des Konstrukts Nutzungsintention (Modell $NI_{(V3\text{-}18)}$)

Modell	χ_r^2 (p-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
$NI_{(V3\text{-}18)}$	83.92 (.000)	5	16.78	**.990**	**.981**	*.073*	[.059, .087]	**.017**

Anmerkungen. N = 4 141 (VERA3 2018); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Auch hinsichtlich der Parameterschätzungen scheinen die Indikatoren geeignet, das latente Konstrukt Nutzungsintention abzubilden (siehe Tabelle 21). Die standardisierten Faktor-

ladungen sind durchweg signifikant ($p < .001$) und liegen alle oberhalb der kritischen Grenze von $\lambda^s \geq .40$ (Berning, 2019) bzw. die meisten Indikatoren oberhalb von $\lambda^s \geq .70$. Lediglich Item NI5 weist eine etwas niedrigere Faktorladung von $\lambda_5^s \geq .657$ auf. Die (Huber-White) Standardfehler (*S. E.* – Standard Error) weisen darüber hinaus keine Auffälligkeiten im Sinne zu großer oder zu kleiner Werte auf, wobei das erste Item (NI1), dessen unstandardisierte Faktorladung auf 1 fixiert wurde, entsprechend keinen Standardfehler aufweist. Die Varianzaufklärung der Items ist größtenteils zufriedenstellend, lediglich die Items NI4 und NI5 liegen etwas unterhalb des Grenzwertes von $R^2 > .50$, jedoch liegt die Varianzaufklärung auch bei diesen Items bei immerhin 49 % bzw. 43 %. Da die beiden Items wichtige Aspekte der Nutzungsintention abbilden und die Bewertung des Gesamtmodellfits und der weiteren Parameter für eine gute Modellpassung sprechen, werden diese beiden Items beibehalten und das Messmodell akzeptiert.

Tabelle 21: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen des Faktors Nutzungsintention basierend auf der konfirmatorischen Faktorenanalyse (Modell $NI_{(V3\text{-}18)}$)

	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
NI1	.873	.000		.762
NI2	.850	.012	77.474	.723
NI3	.832	.012	72.007	.693
NI4	.702	.016	51.260	.492
NI5	.657	.019	42.034	.431

Anmerkungen. $N = 4\,141$ (VERA3 2018); Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Einstellung

Auch bei dem Konstrukt Einstellung erweisen sich alle Interkorrelationen der Indikatorvariablen bei einer zweiseitigen Testung als signifikant ($p < .001$). Die Korrelationskoeffizienten sind dabei mit einem Minimum von $r = .72$ (AE4 $\leftrightarrow$ AE5) alle als hoch einzuschätzen (siehe auch Tabelle 22). Die korrigierte Item-Skala-Korrelation liegt zwischen .85 und .92 und somit deutlich über dem Grenzwert von .50. Auch die DEV überschreitet mit .77 den Grenzwert von .50 deutlich. Der Wert von Cronbachs Alpha ($\alpha = .94$) spricht ebenso für eine sehr gute interne Konsistenz der Skala.

Tabelle 22: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Einstellung (VERA3 2018)

	1.	2.	3.	4.	5.	α	DEV
1. AE1	**.92**						
2. AE2	.78	**.85**					
3. AE3	.85	.78	**.90**			.94	.77
4. AE4	.82	.75	.80	**.87**			
5. AE5	.78	.73	.75	.72	**.85**		

Anmerkungen. $N = 4\,141$ (VERA3 2018); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen; alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test).

Abbildung 13 visualisiert das Messmodell des Konstrukts Einstellung, bei dem wiederum die erste standardisierte Faktorladung auf 1 fixiert wurde.

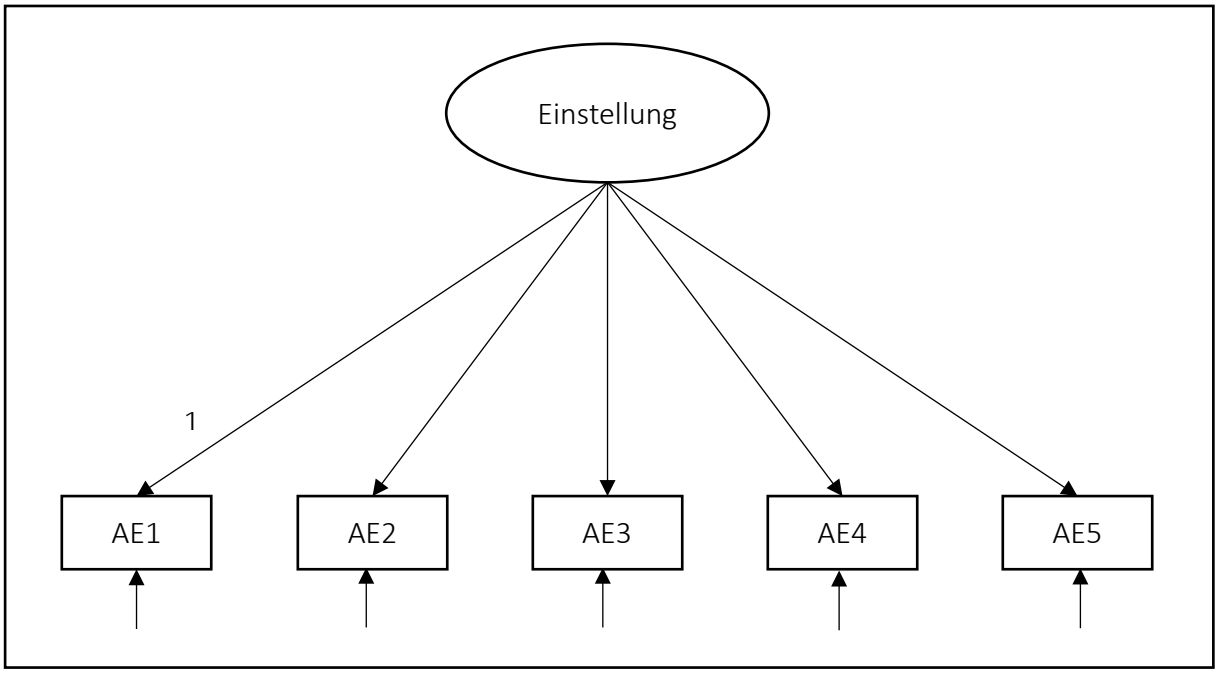


Abbildung 13: Messmodell des Konstrukts Einstellung

Die Schätzung dieses Messmodells (siehe Tabelle 23) weist auf einen insgesamt guten (CFI_r, TLI_r, SRMR) bzw. akzeptablen ($RMSEA_r$) Modellfit hin. Die geschätzten Modellparameter sind in Tabelle 24 dargestellt. Die standardisierten Faktorladungen sind alle signifikant ($p < .001$) und liegen oberhalb des restriktiveren Grenzwertes von $\lambda^s \geq .70$. Die Varianzaufklärung liegt zwischen 69 % und 86 % und somit bei allen Items deutlich über dem

empfohlenen Mindestwert von $R^2 > .50$. Bei der Inspektion der Standardfehler finden sich darüber hinaus keine Auffälligkeiten. Die niedrigste Faktorladung ($\lambda_5^s = .832$) und entsprechend die niedrigste Varianzaufklärung von 69 % weist das Item AE5 auf, welches eine andere Formulierung als die anderen Items hat. Da die Werte jedoch immer noch für eine sehr gute Modellpassung sprechen, wird auch Item AE5 beibehalten und das gesamte Messmodell als passend angenommen.

Tabelle 23: Fitstatistiken der konfirmatorischen Faktorenanalyse des Konstrukts Einstellung (Modell $AE_{(V3\text{-}18)}$)

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
$AE_{(V3\text{-}18)}$	37.59 (.000)	5	7.52	**.997**	**.994**	*.051*	[.036, .066]	**.006**

Anmerkungen. $N = 4\,141$ (VERA3 2018); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Tabelle 24: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen des Faktors Einstellung basierend auf der konfirmatorischen Faktorenanalyse (Modell $AE_{(V3\text{-}18)}$)

	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
AE1	.929	.000		.863
AE2	.851	.011	82.089	.724
AE3	.914	.008	116.553	.835
AE4	.874	.009	100.751	.764
AE5	.832	.011	85.854	.692

Anmerkungen. $N = 4\,141$ (VERA3 2018); Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Aufwand-Nutzen

Tabelle 25 enthält die manifesten Korrelationen der Indikatorvariablen des Konstrukts Aufwand-Nutzen sowie die Trennschärfe (korrigierte Item-Skala-Korrelation) und Cronbachs Alpha. Auch bei diesem Konstrukt korrelieren alle Indikatorvariablen signifikant miteinander (zweiseitige Testung, $p < .001$). Die Stärke der Zusammenhänge ist mit einem minimalen $r = .50$ (AN4 $\leftrightarrow$ AN1r) insgesamt als eher hoch zu bewerten. Die interne Konsistenz kann mit einem Cronbachs Alpha von .91 als sehr gut bewertet werden. Die

Trennschärfekoeffizienten liegen zwischen .77 und .87 und übersteigen somit deutlich den Grenzwert von .50. Die durchschnittliche Varianzaufklärung liegt bei akzeptablen 60 %.

Tabelle 25: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Aufwand-Nutzen (VERA3 2018)

		1.	2.	3.	4.	5.	6.	α	DEV
1.	AN1r [a]	**0.73**							
2.	AN2	0.57	**0.78**						
3.	AN3	0.67	0.79	**0.86**					
4.	AN4	0.50	0.52	0.57	**0.77**			.91	.60
5.	AN5	0.56	0.60	0.64	0.76	**0.85**			
6.	AN6	0.68	0.64	0.73	0.65	0.76	**0.87**		

Anmerkungen. $N = 4\,141$ (VERA3 2018); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen; [a] Das negativ gepolte Item AN1 wurde für die Berechnungen umgepolt (= AN1r); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test).

Die Schätzung des Messmodells dieses Konstrukts erfolgte in mehreren Schritten. In Tabelle 26 sind die Fitstatistiken der zu diesem Zweck geschätzten Modelle aufgeführt. Zunächst wurde ein Modell mit vollständig unkorrelierten Residuen geschätzt (Modell $AN1_{(V3\text{-}18)}$). Dieses weist jedoch, abgesehen von einem SRMR = .048, einen schlechten Modellfit auf.

Tabelle 26: Fitstatistiken der konfirmatorischen Faktorenanalysen des Konstrukts Aufwand-Nutzen (Modelle $AN1_{(V3\text{-}18)} – AN3_{(V3\text{-}18)}$)

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	AIC
$AN1_{(V3\text{-}18)}$	1 292.90 (.000)	9	143.66	.893	.822	.222	[.212, .232]	**.048**	48 810
$AN2_{(V3\text{-}18)}$	293.31 (.000)	6	48.89	**.981**	*.952*	.115	[.104, .127]	**.023**	47 325
$AN3_{(V3\text{-}18)}$	77.83 (.000)	5	15.57	**.995**	**.986**	*.062*	[.050, .074]	**.012**	**47 079**

Anmerkungen. $N = 4\,141$ (VERA3 2018); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben;
Modell $AN1_{(V3\text{-}18)}$: 1-Faktor-Lösung.
Modell $AN2_{(V3\text{-}18)}$: Residualkorrelation zwischen den Items AN4, AN5 und AN6.
Modell $AN3_{(V3\text{-}18)}$: Zusätzliche Residualkorrelation zwischen den Items AN2 und AN3.

Die bei der Modellschätzung mit lavaan ausgegebenen Modifikationsindizes, also Hinweise zur Verbesserung des Modellfits, schlagen vor, zwischen bestimmten Indikatorvariablen Residualkorrelationen zuzulassen. Da die Variablen AN4, AN5 und AN6 eine größtenteils gleichlautende Formulierung („Im Verhältnis zu den gewonnenen Erkenntnissen empfinde ich den Zeitaufwand der ... Vorbereitung/Durchführung/Auswertung von VERA als ...“) einschließlich einer anderen Skala (1 = *nicht angemessen* bis 4 = *angemessen*) aufweisen, liegt die Vermutung nahe, dass hierin eine gemeinsame Ursache der Messfehler begründet liegt. Deshalb wurden in Modell $AN2_{(V3\text{-}18)}$ die entsprechenden Residualkorrelationen spezifiziert, was zu einer deutlichen Modellverbesserung führt. Der Wert des $RMSEA_r$ erfüllt jedoch mit .115 noch nicht die Kriterien eines ausreichend gut fittenden Modells. Daher wurde mit Blick auf die Modifikationsindizes und die Itemformulierungen eine weitere Residualkorrelation spezifiziert. Die ersten drei Items des Konstrukts AN1 bis AN3 wurden alle mit der Skala 1 = *stimme überhaupt nicht zu* bis 4 = *stimme voll und ganz zu* gemessen, das Item AN1 weist jedoch eine inverse Formulierung auf und hebt sich dadurch von den anderen beiden Items ab. Um diesen möglichen Einfluss auf das Antwortverhalten abzubilden, wurde in Modell $AN3_{(V3\text{-}18)}$ zusätzlich die Residualkorrelation zwischen den Items AN2 und AN3 zugelassen.

Das finale Messmodell $AN3_{(V3\text{-}18)}$, das in Abbildung 14 dargestellt ist, weist schließlich einen insgesamt zufriedenstellenden Modellfit auf. CFI_r, TLI_r und SRMR zeigen gute Kennwerte auf, der $RMSEA_r$ zumindest akzeptable. Zusätzlich kennzeichnet der AIC Modell $AN3_{(V3\text{-}18)}$ in Relation zu der Zahl seiner zu schätzenden Parameter als effizienter im Vergleich zu den anderen beiden Modellen.

Die Parameterschätzungen von Modell $AN3_{(V3\text{-}18)}$ sind in Tabelle 27 dokumentiert. Die standardisierten Faktorladungen aller Items sind signifikant ($p < .001$) und liegen mit Ausnahme des Items AN4 alle oberhalb $\lambda^s \geq .70$. Jedoch liegt die Faktorladung dieses Items mit .662 immer noch deutlich über der kritischen Grenze von .40. Die robusten Standardfehler sind auch hier unauffällig. Die Varianzaufklärung der Indikatorvariablen liegt zwischen 44 % und 74 %, wobei wiederum nur das Item AN4 etwas unterhalb des Grenzwertes von $R^2 > .50$ liegt. Aufgrund der davon abgesehen guten Passung des Modells erscheint diese Abweichung tolerierbar und das Modell $AN3_{(V3\text{-}18)}$ geeignet, die empirischen Daten abzubilden.

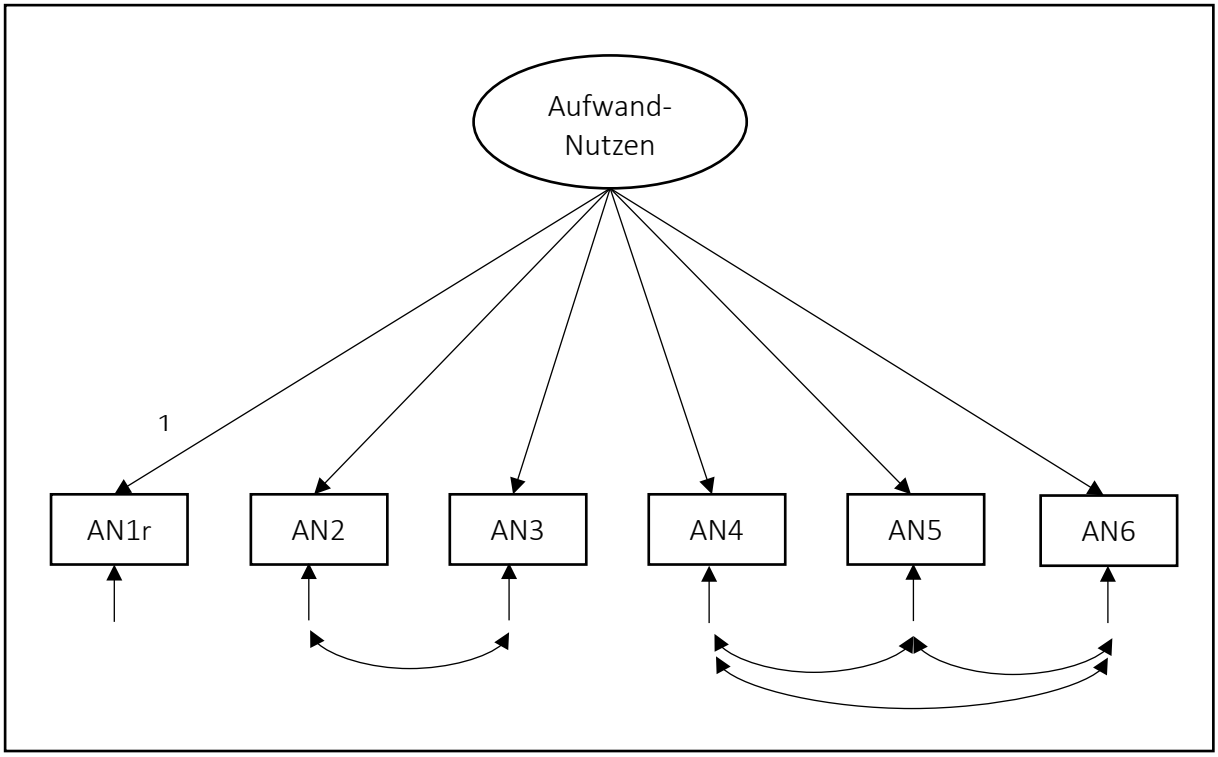


Abbildung 14: Messmodell des Konstrukts Aufwand-Nutzen (Modell AN3$_{(V3\text{-}18)}$)

Tabelle 27: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen des Faktors Aufwand-Nutzen basierend auf der konfirmatorischen Faktorenanalyse (Modell AN3$_{(V3\text{-}18)}$)

	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
AN1r	.776	.000		.602
AN2	.747	.021	43.349	.558
AN3	.851	.020	53.373	.724
AN4	.662	.025	38.393	.438
AN5	.747	.025	43.750	.558
AN6	.858	.023	56.688	.736

Anmerkungen. N = 4 141 (VERA3 2018); Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Die in Modell AN3$_{(V3\text{-}18)}$ spezifizierten Residualkorrelationen liegen zwischen .21. und .54 und fallen durchweg signifikant aus ($p < .001$) (siehe Tabelle 28).

Tabelle 28: Spezifizierte Residualkorrelationen der Indikatorvariablen der konfirmatorischen Faktorenanalyse (Modell AN3$_{(V3\text{-}18)}$) des Konstrukts Aufwand-Nutzen (vollständig standardisierte Lösung)

	AN3	AN5	AN6
AN2	.43	-	-
AN3	-	-	-
AN4	-	.54	.21
AN5	-	-	.34

Anmerkungen. Alle Koeffizienten erweisen sich als signifikant ($p < .001$).

Nützlichkeit und zeitliche Belastung

Die Konstrukte Nützlichkeit und zeitliche Belastung wurden mit einem gemeinsamen Messmodell geschätzt, da das Konstrukt zeitliche Belastung mit drei Indikatoren ohne weitere Restriktionen nur gerade identifiziert wäre ($df = 0$) (Brown, 2015) und solche Restriktionen, wie bspw. gleichgesetzte Faktorladungen (tau-Äquivalenz, vgl. Steyer & Eid, 1993), inhaltlich wenig plausibel wären. Zunächst wurden jedoch Interkorrelationen, Trennschärfe und interne Konsistenz je Konstrukt untersucht. Die entsprechenden Kennwerte des Konstrukts Nützlichkeit sind in Tabelle 29 dargestellt. Die manifesten Korrelationen der Indikatorvariablen erweisen sich als signifikant ($p < .001$, zweiseitige Testung). Die Höhe der Korrelationen liegt zwischen $r = .54$ und $r = .67$. Die korrigierte Item-Skala-Korrelation und Cronbachs Alpha zeigen ebenfalls zufriedenstellende Werte. Die Trennschärfe liegt zwischen .73 und .84, Cronbachs Alpha bei .89. Auch die durchschnittliche Varianzaufklärung von 62 % verdeutlicht die Konstruktreliabilität der Skala.

Tabelle 29: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Nützlichkeit (VERA3 2018)

		1.	2.	3.	4.	5.	α	DEV
1.	WN1	**.79**						
2.	WN2	.60	**.77**					
3.	WN3	.67	.67	**.84**			.89	.62
4.	WN4	.59	.54	.63	**.73**			
5.	WN5	.62	.63	.65	.57	**.79**		

Anmerkungen. $N = 4141$ (VERA3 2018); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen; alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test).

Die analogen Kennwerte des Konstrukts zeitliche Belastung sind in Tabelle 30 aufgeführt. Auch hier sind alle Interkorrelationen zwischen den Indikatoritems bei einer zweiseitigen Testung signifikant ($p < .001$). Die Höhe der Korrelationskoeffizienten spricht für einen moderaten ($r = .34$ bzw $r = .43$.) bzw. eher geringen Zusammenhang ($r = .19$). zwischen den Items. Die Werte der korrigierten Item-Skala-Korrelation liegen zwischen .51 und .65 und somit alle über dem Grenzwert von .50. Cronbachs Alpha liegt mit $\alpha = .57$ zwar unterhalb des angestrebten Wertes von .70, kann jedoch bei einer Skala mit nur drei Indikatorvariablen noch als akzeptabel angesehen werden (Zinnbauer & Eberl, 2004). Auch die DEV fällt mit .32 unter den Schwellenwert von .50, was für eine nicht optimale Repräsentation der Skala durch die gewählten Items spricht. Mögliche Ursachen hierfür und entsprechende Konsequenzen werden später in diesem Abschnitt weiter erläutert.

Tabelle 30: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Zeitliche Belastung (VERA3 2018)

	1.	2.	3.	α	DEV
1. ZB1	**.51**				
2. ZB2	.34	**.65**		.57	.32
3. ZB3	.19	.43	**.59**		

Anmerkungen. $N = 4\,141$ (VERA3 2018); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen; alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test).

Das Ergebnis der Modellschätzung ist in Tabelle 31 aufgeführt. Alle Fitstatistiken zeigen einen guten Modellfit an.

Tabelle 31: Fitstatistiken der konfirmatorischen Faktorenanalyse der Konstrukte Zeitliche Belastung und Nützlichkeit (Modell WN/ZB$_{(V3\text{-}18)}$)

Modell	χ_r^2 (p-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
WN/ZB$_{(V3\text{-}18)}$	123.64 (.000)	19	6.51	**.990**	**.985**	**.039**	[.033, .046]	**.030**

Anmerkungen. $N = 4\,141$ (VERA3 2018); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Das gemeinsame Messmodell der beiden Konstrukte als zwei korrelierte Faktoren ist in Abbildung 15 dargestellt. Wie bei den anderen Messmodellen wurde jeweils die erste standardisierte Faktorladung auf 1 fixiert.

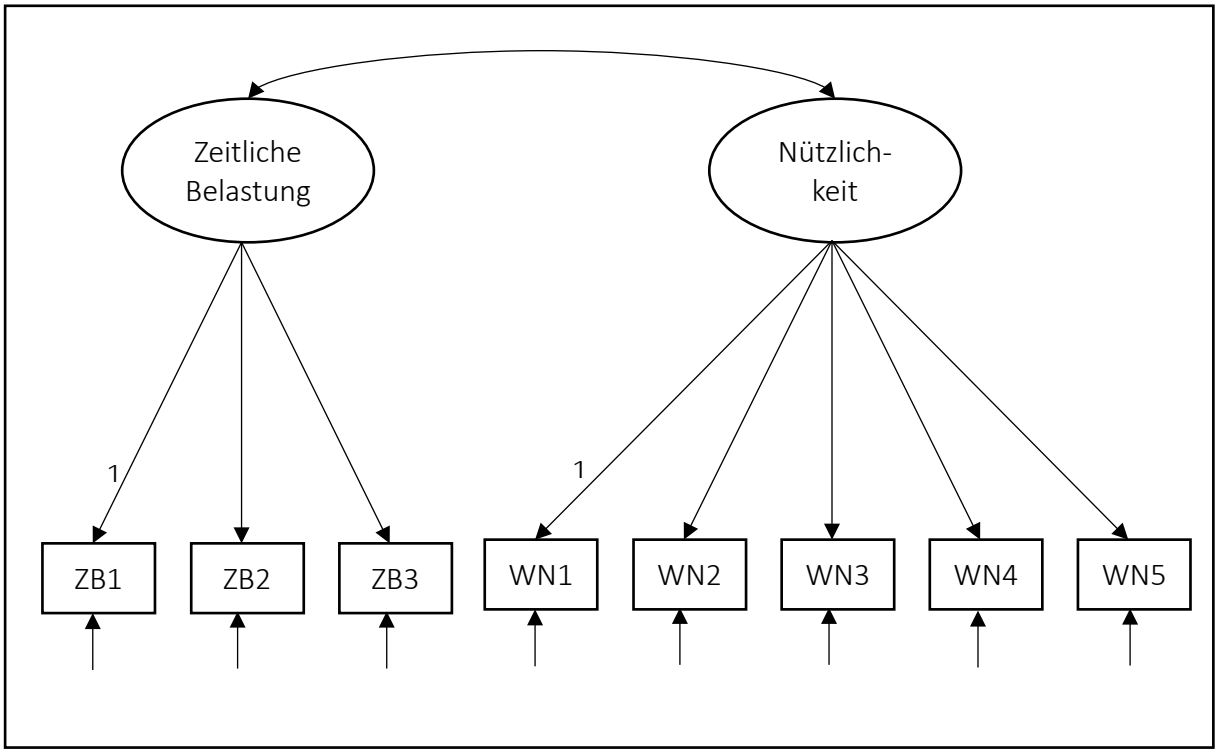


Abbildung 15: Messmodell der Konstrukte Zeitliche Belastung und Nützlichkeit

Die Korrelationen zwischen den Konstrukten erweist sich bei zweiseitiger Testung als signifikant ($p < .001$), der Korrelationskoeffizient liegt bei $r = -.27$. Hinsichtlich der weiteren Parameterschätzungen (siehe Tabelle 32) zeichnet sich bei dem Konstrukt Nützlichkeit ein konsistent zufriedenstellendes Bild ab, die Indikatoren erscheinen durchweg gut geeignet, das latente Konstrukt Nützlichkeit abzubilden. Alle standardisierten Faktorladungen sind signifikant ($p < .001$), liegen zwischen .730 und .851 und somit über dem restriktiveren Grenzwert von $\lambda^s \geq .70$. Die Standardfehler sind alle unauffällig, und die Varianzaufklärung liegt zwischen 53 % und 73 % und somit oberhalb der kritischen Grenze von $R^2 > .50$. Die Bewertung der Modellparameter des Konstrukts zeitliche Belastung ist im Vergleich weniger eindeutig. Das Item ZB2 hat mit $\lambda^s \geq .803$ eine sehr hohe Faktorladung und mit 65 % auch eine gute Varianzaufklärung. Die beiden anderen Items ZB1 und ZB3 erreichen bzw. überschreiten nur die Minimalanforderungen an einen akzeptablen Indikator, einer Faktorladung von $\lambda^s \geq .40$ (Berning, 2019) bzw. $\lambda^s \geq .50$ (Urban & Mayerl, 2014). Die Varianzaufklärung liegt mit 16 % bzw. 29 % deutlich unter dem angestrebten Richtwert von 50 %. Die Standardfehler sind zwar

im Verhältnis zu dem Konstrukt Nützlichkeit vergleichsweise hoch, jedoch ähnlich groß für die Indikatoren innerhalb des Konstrukts.

Tabelle 32: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen der Faktoren Nützlichkeit und Zeitliche Belastung basierend auf der konfirmatorischen Faktorenanalyse (Modell WN/ZB$_{(V3\text{-}18)}$)

	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
WN1 [a]	.784	.000		.614
WN2 [a]	.780	.020	52.292	.608
WN3 [a]	.851	.020	59.085	.724
WN4 [a]	.730	.021	48.332	.533
WN5 [a]	.783	.021	51.149	.614
ZB1 [b]	.400	.000		.160
ZB2 [b]	.803	.096	15.970	.646
ZB3 [b]	.536	.078	14.876	.287

Anmerkungen. N = 4 141 (VERA3 2018). [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Insgesamt erweist sich das Konstrukt zeitliche Belastung als problematisch, da die beiden Indikatoren ZB1 und ZB3 das Konstrukt basierend auf einer Beurteilung der geschätzten Modellparameter nicht optimal widerspiegeln. Gerade das Item ZB1 stellt sich bereits bei der Betrachtung der Interkorrelationen als weniger verbunden mit den beiden anderen Indikatoren heraus. Zusätzlich gibt auch die Betrachtung der Itemmittelwerte, durch einen vergleichsweise niedrigen Mittelwert und eine größere Standardabweichung, Hinweise auf eine Abweichung dieses Items (siehe Tabelle 18, S. 164). Auf der Suche nach möglichen Ursachen muss der Inhalt der Items berücksichtigt werden. Jedes Item des Konstrukts erfasst unterschiedliche Aspekte der zeitlichen Belastung im Hinblick auf die einzelnen Arbeitsschritte der Vergleichsarbeiten („Wie bewerten Sie den Zeitaufwand für die Vorbereitung (ZB1)/Durchführung (ZB2)/Auswertung (ZB3)?“). Streng genommen sind diese Schritte als Komplemente, nicht als Substitute zu betrachten, wie eigentlich von reflektiven Indikatoren gefordert. Die Wegnahme eines Indikators würde dazu führen, dass ein Aspekt des Zeitaufwandes fehlen würde. Dies könnte ein Hinweis darauf sein, dass im Falle dieser Skala eine formative Modellierung besser geeignet wäre (siehe auch Diskussion, Kapitel 6.1.2, Skalenvalidierung).

Zwar liefern die niedrigen Korrelationen der Items und die nicht optimale Passung einer reflektiven Modellierung Hinweise darauf, dass es sich evtl. eher um ein formatives Konstrukt

handeln könnte, jedoch keine eindeutigen Beweise. Bei der Interpretation gilt des Weiteren zu beachten, dass eine geringe interne Konsistenz etc. weder eine notwendige noch eine hinreichende Voraussetzung einer formativen Modellierung darstellt (Edwards, 2011). Um diesbezüglich eine klare Aussage treffen zu können, müssten detaillierte Analysen durchgeführt werden, auf die aus den in Kapitel 4.3.2 genannten pragmatischen Gründen jedoch verzichtet wurde, weil auch diese nicht immer zu einem eindeutigen Ergebnis führen. Der Versuch einer formativen Modellierung würde aufgrund der Gefahr, weitere uneindeutige Ergebnisse zu erhalten, den Aufwand einer alternativen Modellierung nicht rechtfertigen. Zusätzlich ist die Entscheidung für oder gegen die eine oder andere Spezifikationsart immer im Voraus zu treffen und sollte keinesfalls a posteriori aufgrund einer schlechten Passung einer reflektiven Modellierung angepasst werden (Fluck, 2020a). Da eine ausführliche Diskussion und Untersuchung verschiedener Spezifikationsarten im Rahmen dieser Arbeit nicht leistbar ist, wird der Faktor zeitliche Belastung trotz einer nicht optimalen Passung, aber aufgrund des insgesamt zufriedenstellenden Modellfits mit einer reflektiven Modellierung beibehalten und das dargestellte Messmodell $WN/ZB_{(V3\text{-}18)}$ unter Vorbehalt akzeptiert.

CFA mit allen Konstrukten

In einem letzten Schritt der Skalenanalyse wurden alle Konstrukte in einem Messmodell geschätzt. Das Messmodell wurde zunächst analog zu den zuvor separat geschätzten konfirmatorischen Faktorenanalysen spezifiziert und Interkorrelationen zwischen allen fünf latenten Faktoren zugelassen (siehe Tabelle 33, Modell $1_{CFA(V3\text{-}18)\_a}$). Der TLI_r liegt mit .942 unterhalb des Cutoff-Wertes für eine akzeptable Modellpassung von .950. $RMSEA_r$ und CFI_r sind mit .062 bzw. .950 als (noch) akzeptabel zu bewerten, der SRMR mit .037 sogar als gut. Insgesamt betrachtet, erscheint der Modellfit jedoch noch verbesserungsfähig.

Für eine Verbesserung des Modellfits wurden mit Blick auf die Itemformulierungen zusätzlich Residualkorrelationen zwischen Items der Faktoren zeitliche Belastung und Aufwand-Nutzen spezifiziert (AN4 ↔ ZB1, AN5 ↔ ZB2, AN6 ↔ ZB3). Die Items zielen auf verschiedene Aspekte des jeweils gleichen Bewertungskriteriums ab und weisen entsprechend teils gleiche Formulierungen auf, sodass dies als Ursache für das Auftreten einer gemeinsamen Messfehlervarianz gesehen werden kann. Die betroffenen Items des Konstrukts Aufwand-Nutzen (AN4, AN5, AN6) haben folgende Formulierung: „Im Verhältnis zu den gewonnenen Erkenntnissen empfinde ich den Zeitaufwand der…Vorbereitung/Durchführung/Auswertung von VERA

als…“. Die Items des Konstrukts zeitliche Belastung (ZB1, ZB2, ZB3) lauten: „Wie bewerten Sie den Zeitaufwand für die Vorbereitung/Durchführung/Auswertung?“.

Tabelle 33: Fitstatistiken der konfirmatorischen Faktorenanalyse mit allen Konstrukten (Modell $1_{CFA(V3\text{-}18)\_a}$ und $1_{CFA(V3\text{-}18)\_b}$)

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	AIC
$1_{CFA(V3\text{-}18)\_a}$	3 502.50 (.000)	238	14.72	*.950*	.942	*.062*	[.060, .064]	**.037**	17 7525
$1_{CFA(V3\text{-}18)\_b}$	1 952.02 (.000)	235	8.31	**.974**	*.969*	**.045**	[.043, .047]	**.033**	**17 5746**

Anmerkungen. $N = 4\,141$ (VERA3 2018); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.
Modell $1_{CFA(V3\text{-}18)\_a}$: 5 Faktoren ohne Residualkorrelationen zwischen latenten Faktoren.
Modell $1_{CFA(V3\text{-}18)\_b}$: 5 Faktoren mit Residualkorrelationen zwischen latenten Faktoren: AN4 ↔ ZB1, AN5 ↔ ZB2, AN6 ↔ ZB3.

Diese Anpassung des Messmodells führt zu einer deutlichen Verbesserung des Modellfits (siehe Tabelle 33, Modell $1_{CFA(V3\text{-}18)\_b}$). Die Fitindikatoren weisen einen guten bzw. akzeptablen Modellfit aus: $CFI_r = .974$, $RMSEA_r = .045$, SRMR = .033 sowie $TLI_r = .969$. Darüber hinaus weist auch der AIC darauf hin, Modell $1_{CFA(V3\text{-}18)\_b}$ im Vergleich zu Modell $1_{CFA(V3\text{-}18)\_a}$ zu bevorzugen. Die Parameterschätzungen von Modell $1_{CFA(V3\text{-}18)\_b}$ sind vergleichbar mit denen der separaten Modellschätzungen und in Anhang B zu finden. Die in Modell $1_{CFA(V3\text{-}18)\_b}$ spezifizierten Residualkorrelationen sind dabei signifikant ($p < .001$) und liegen bei -.39 (AN4 ↔ ZB1), -.22 (AN5 ↔ ZB2) und -.41 (AN6 ↔ ZB3).

Tabelle 34 enthält die Korrelationskoeffizienten zwischen den latenten Konstrukten. Alle dargestellten Korrelationen erweisen sich als signifikant ($p < .001$, zweiseitige Testung), und der Betrag der Korrelationskoeffizienten liegt zwischen .24 und .92. Erwartungsgemäß korreliert der Faktor zeitliche Belastung aufgrund seiner inversen Polung negativ mit den übrigen Faktoren. Die Korrelationskoeffizienten liegen zwischen -.24 und -.49 und somit im mittleren Bereich. Die Interkorrelationen zwischen den übrigen Konstrukten sind dagegen höher und liegen zwischen .72 und .92. Insgesamt liegen die meisten Korrelationen zwischen den latenten Konstrukten unter dem Grenzwert von .85 (Brown, 2015) bzw. .90 (Kline, 2011), wonach die latenten Faktoren unterschiedliche, klar trennbare Konstrukte repräsentieren. Die Korrelation zwischen den Konstrukten Nützlichkeit und Aufwand-Nutzen beträgt zwar .85, nach den Kriterien von Kline (2011) kann dieser Wert aber noch als akzeptabel angesehen werden. Die

Korrelation zwischen den Faktoren Aufwand-Nutzen und Einstellung übersteigt hingegen mit $r = .92$ auch den von Kline (2011) postulierten weniger restriktiven Grenzwert von $r = .90$. Dies könnte ein Indiz dafür sein, dass die Proband*innen die beiden Konstrukte nicht wie theoretisch erwartet differenzierten. Aus der Formulierung der Indikatorvariablen der beiden Konstrukte ergeben sich diesbezüglich jedoch keine Hinweise. Brown (2015) und Kline (2011) schlagen beim Auftreten einer zu hohen Interkorrelation zwischen zwei latenten Konstrukten vor, die Indikatorvariablen beider Konstrukte zu einem Faktor zusammenzufassen und zu prüfen, ob sich der Modellfit verschlechtert.

Tabelle 34: Korrelationen der latenten Faktoren (Modell $1_{CFA(V3\text{-}18)\_b}$) (vollständig standardisierte Lösung) sowie manifeste Skalenmittelwerte und Standardabweichungen

		1.	2.	3.	4.	5.	*M*	*SD*	*n*
1.	Nutzungsintention [a]						2.37	0.69	3 862
2.	Einstellung [a]	.73					2.38	0.82	3 975
3.	Aufwand-Nutzen [a]	.72	.92				2.50	0.78	3 721
4.	Zeitliche Belastung [b]	-.24	-.37	-.50			3.06	0.62	4 063
5.	Nützlichkeit [a]	.79	.80	.85	-.32		2.49	0.74	3 832

Anmerkungen. $N = 4\,141$ (VERA3 2018). [a] 4-stufige positiv gepolte Antwortskala; [b] 5-stufige negativ gepolte Antwortskala; alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test).

Im vorliegenden Fall führt die Schätzung eines Modells mit einer gemeinsamen Modellierung der Konstrukte Einstellung und Aufwand-Nutzen als ein Faktor im Vergleich zu Modell $1_{CFA(V3\text{-}18)\_b}$ zu einer Verschlechterung des Modellfits: $CFI_r = .960$, $TLI_r = .954$, $RMSEA_r = .055$ sowie $SRMR = .039$. Da eine solche Modellierung der Konstrukte als gemeinsamer Faktor auch inhaltlich und hinsichtlich der Formulierung der Items nicht plausibel erscheint, wird das ursprüngliche Modell $1_{CFA(V3\text{-}18)\_b}$ zunächst in einem ersten Schritt beibehalten. In einem weiteren Arbeitsschritt werden jedoch unter Berücksichtigung dieses starken Zusammenhangs Modellanpassungen vorgenommen (siehe Kapitel 5.2).

Neben den latenten Faktorkorrelationen sind in Tabelle 34 die manifesten Skalenmittelwerte und Standardabweichungen wiedergegeben. Die letzte Spalte gibt zudem die der jeweiligen Skalenbildung zugrundeliegende Stichprobengröße an, da nur jeweils vollständige Fälle für die

Mittelwertbildung berücksichtigt wurden. Unter den mit einer 4-stufigen Skala gemessenen Konstrukten weist der Faktor Aufwand-Nutzen mit 2.50 (*SD* = 0.78) den höchsten Mittelwert auf, welcher exakt dem theoretischen Skalenmittel entspricht. Die Bewertung der wahrgenommenen Nützlichkeit liegt mit *M* = 2.49 (*SD* = 0.74) knapp unter dem theoretischen Mittel. Etwas niedriger fallen die Mittelwerte der Nutzungsintention (*M* = 2.37, *SD* = 0.69) und der Einstellung (*M* = 2.38, *SD* = 0.82) aus. Bei diesen Konstrukten ist somit insgesamt eine eher negative Antworttendenz erkennbar. Auch die Bewertung der zeitlichen Belastung weist eine insgesamt leicht negative Tendenz auf, der Mittelwert dieses Konstrukts liegt mit *M* = 3.06 (*SD* = 0.62) zwar knapp über dem theoretischen Mittelwert, jedoch muss die inverse Polung der Items berücksichtigt werden.

Als Gesamtfazit dieses Kapitels kann festgehalten werden, dass die postulierten Messmodelle trotz einzelner Defizite, die bei der Ergebnisinterpretation berücksichtigt werden müssen, insgesamt gut zu den Daten passen und für eine Weiterarbeit basierend auf dem erläuterten Modell $1_{CFA(V3\text{-}18)\_b}$ geeignet sind. Die Bewertung der einzelnen Konstrukte durch die Lehrkräfte weist im Gesamten eine negative Tendenz auf, wobei die Zusammenhänge der einzelnen Bewertungsaspekte nun im folgenden Kapitel anhand der spezifizierten Strukturmodelle untersucht werden.

5.1.3. Untersuchung der Kausalbeziehungen (VERA3 2018)

Nach der Schätzung eines validen Messmodells erfolgte die Analyse des Strukturmodells in mehreren Schritten: Zunächst wurde in Anlehnung an das ursprüngliche TAM (F. D. Davis, 1986) ein Modell geschätzt, in dem keine direkte Beziehung zwischen Nützlichkeit und Nutzungsintention spezifiziert wurde (siehe Modell $1_{(V3\text{-}18)}$). Dem liegt die theoretische Annahme zugrunde, dass Wahrnehmungen bzw. Überzeugungen – hier Nützlichkeit, zeitliche Belastung und Aufwand-Nutzen-Bewertung – nicht direkt verhaltenswirksam werden bzw. die Verhaltensintention beeinflussen, sondern nur über die Einstellungsbildung wirken (Eagly & Chaiken, 1993; Fishbein & Ajzen, 1975). Da jedoch in Folgearbeiten die direkte Beziehung zwischen Nützlichkeit und Nutzung(sintention) vielfach empirisch belegt worden ist (Chismar & Wiley-Patton, 2003; siehe bspw. F. D. Davis, 1993; Ho Cheong & Park, 2005; Zhao et al., 2018), wurde dieser direkte Pfad im zweiten Schritt in Modell $1_{dP(V3\text{-}18)}$ spezifiziert. Um u. a. der Problematik der starken latenten Korrelation zwischen den Konstrukten Aufwand-Nutzen und

Einstellung zu begegnen (siehe auch Kapitel 5.1.2, Tabelle 34), wurde das Modell erneut angepasst und das Konstrukt Aufwand-Nutzen aus der Modellschätzung entfernt (siehe Kapitel 5.2).

Modell 1$_{(V3\text{-}18)}$

Abbildung 16 zeigt Modell 1$_{(V3\text{-}18)}$ ohne den direkten Pfad zwischen Nützlichkeit und Nutzungsintention. Dargestellt sind die standardisierten Faktorladungen, Varianzaufklärung je Konstrukt (R^2) und die Fitstatistiken des Modells. Die Gütekriterien sprechen für einen akzeptablen Modellfit. Die Werte des CFI_r, und TLI_r liegen mit .966 und .960 im akzeptablen Bereich, ebenso wie der $RMSEA_r$ = .051. Der SRMR weist mit .046 sogar einen guten Wert auf. Das Messmodell wurde entsprechend der CFA mit allen Faktoren im vorangegangenen Abschnitt spezifiziert. Die zugehörigen Parameterschätzungen wie Faktorladungen und Residualkorrelationen sind daher mit denen der Faktorenanalyse vergleichbar und in Anhang C dargestellt.

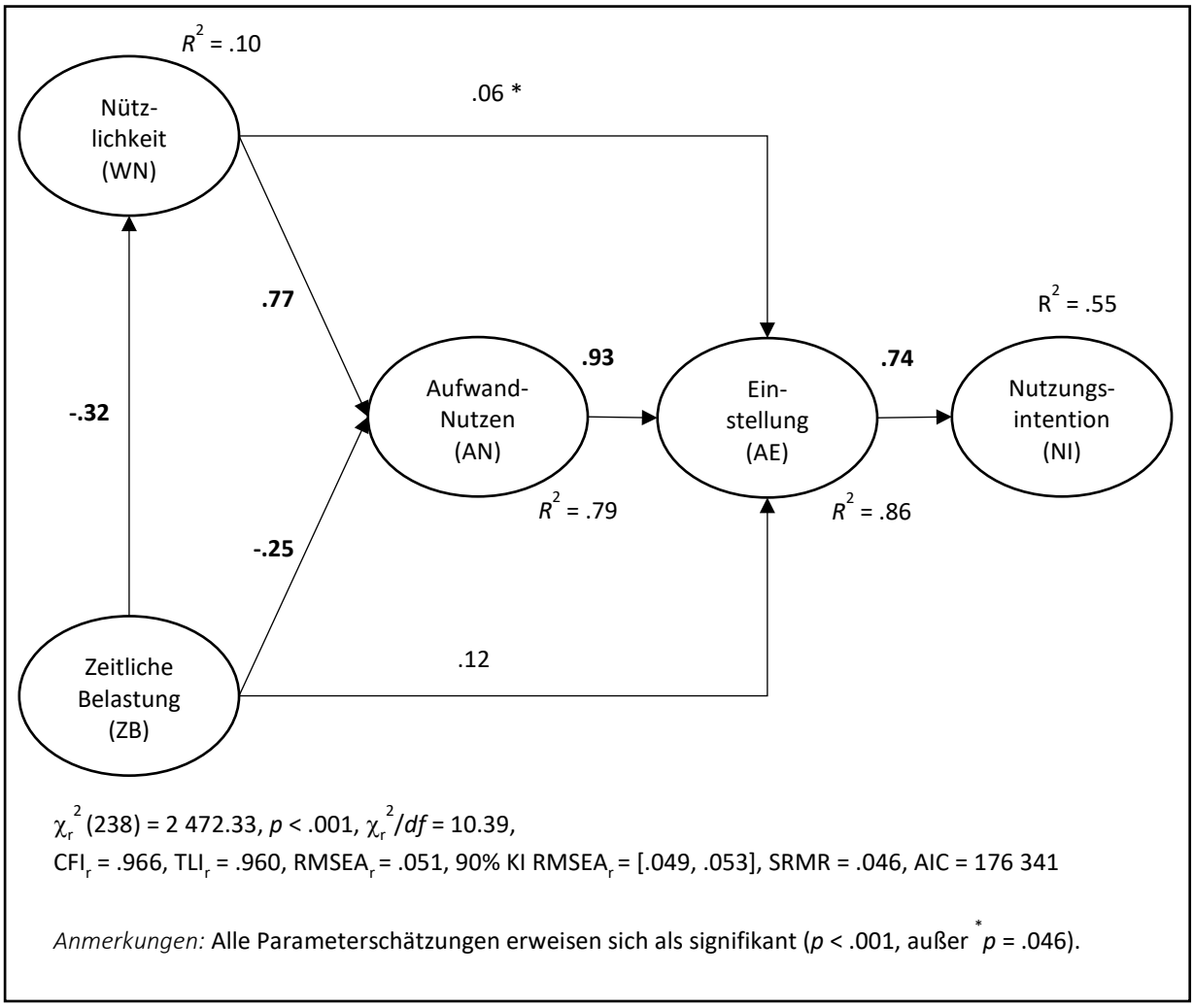


Abbildung 16: Modell 1$_{(V3\text{-}18)}$ – Fitstatistiken und Parameterschätzungen (vollständig standardisierte Lösung).

Mit Ausnahme des direkten Pfades der wahrgenommenen Nützlichkeit auf die Einstellung, werden alle im Modell geschätzten Pfadkoeffizienten auf dem 1 % Niveau signifikant. Diese Beziehung wird nur auf dem 5 % Niveau signifikant und markiert mit β = .06 eine nur sehr schwache Beziehung. Alle Beziehungen, abgesehen von dem Pfad zwischen zeitlicher Belastung und Einstellung, weisen dabei das erwartete Vorzeichen auf. Die exogene Variable zeitliche Belastung wirkt negativ sowohl auf die wahrgenommene Nützlichkeit (γ = -.32) als auch auf das wahrgenommene Aufwand-Nutzen-Verhältnis (γ = -.25). Entgegen den Erwartungen wirkt eine hohe zeitliche Belastung jedoch positiv auf die Einstellung (γ = .12). Mögliche Erklärungen hierfür werden in Kapitel 6.1.2, S. 230ff. aufgegriffen.

Der Einfluss der zeitlichen Belastung erklärt 10 % der Varianz des Konstrukts Nützlichkeit (R^2 = .10). Die wahrgenommene Nützlichkeit wiederum wirkt verstärkend auf die Aufwand-Nutzen-Bewertung (β = .77) und erklärt zusammen mit der zeitlichen Belastung 79 % der Varianz dieses Konstrukts (R^2 = .79). Die Aufwand-Nutzen-Abwägung beeinflusst die Einstellung mit β = .93 extrem stark. Zusammen mit dem positiven Einfluss der zeitlichen Belastung und dem marginalen direkten Effekt der Nützlichkeit (β = .06) werden insgesamt 89 % der Varianz des Konstrukts Einstellung durch das Modell erklärt (R^2 = .89). Schlussendlich wirkt die Einstellung positiv auf die Nutzungsintention (β = .74) und führt zu einer Varianzaufklärung des Konstrukts von 55 % (R^2 = .55).

Modell $1_{dP(V3\text{-}18)}$

Eine Anpassung des Modells durch die Spezifikation eines direkten Pfades zwischen wahrgenommener Nützlichkeit und Nutzungsintention (Modell $1_{dP(V3\text{-}18)}$, siehe Abbildung 17) führt zu einer deutlichen Verbesserung der Modellgüte. Die Fitstatistiken weisen im Vergleich zu Modell $1_{(V3\text{-}18)}$ allesamt verbesserte Werte auf: CFI_r = .974, TLI_r = .969, $RMSEA_r$ = .045, SRMR = .033. Auch der Wert des AIC unterstreicht die Effizienz des Modells $1_{dP(V3\text{-}18)}$ verglichen mit Modell $1_{(V3\text{-}18)}$ ($AIC_{Modell1(V3\text{-}18)}$ = 176 341, $AIC_{Modell1dP(V3\text{-}18)}$ = 175 752). Die Parameterschätzungen des Messmodells sind ebenso wie für Modell $1_{(V3\text{-}18)}$ in Anhang C dargestellt.

Ein Großteil der Parameterschätzungen des Strukturmodells zeigt keine bzw. nur marginale Veränderungen der Schätzwerte. Die Pfadkoeffizienten der exogenen Variablen zeitliche Belastung bspw. verändern sich kaum: Die Wirkung auf die Einstellung bleibt mit γ = .12 unverändert, der Effekt auf die Aufwand-Nutzen-Bewertung (γ = -.25) sowie die Nützlichkeit

(γ = -.32) variieren im Vergleich zu Modell $1_{(V3\text{-}18)}$ jeweils um .01. Alle drei Effekte sind auf dem 1 % Niveau signifikant. Die Varianzaufklärung der wahrgenommenen Nützlichkeit durch den Einfluss der zeitlichen Belastung beträgt auch hier 10 %. Auch die Wirkung der Nützlichkeit auf das Konstrukt Aufwand-Nutzen entspricht mit β = .77 (p < .001) dem Effekt in Modell $1_{(V3\text{-}18)}$. Die Varianzaufklärung dieses Konstrukts durch Nützlichkeit und zeitliche Belastung liegt bei 78 % und ist somit ebenfalls mit Modell $1_{(V3\text{-}18)}$ vergleichbar. Der bereits im ersten Modell sehr schwache direkte Effekt der Nützlichkeit auf die Einstellung wird in diesem Modell insignifikant. Die Varianzaufklärung der Einstellung, die mit β = .96 (p < .001) ähnlich stark wie in Modell $1_{(V3\text{-}18)}$ durch die Aufwand-Nutzen-Abwägung beeinflusst wird, liegt bei 85 %, also kaum verändert im Vergleich zum ersten Modell.

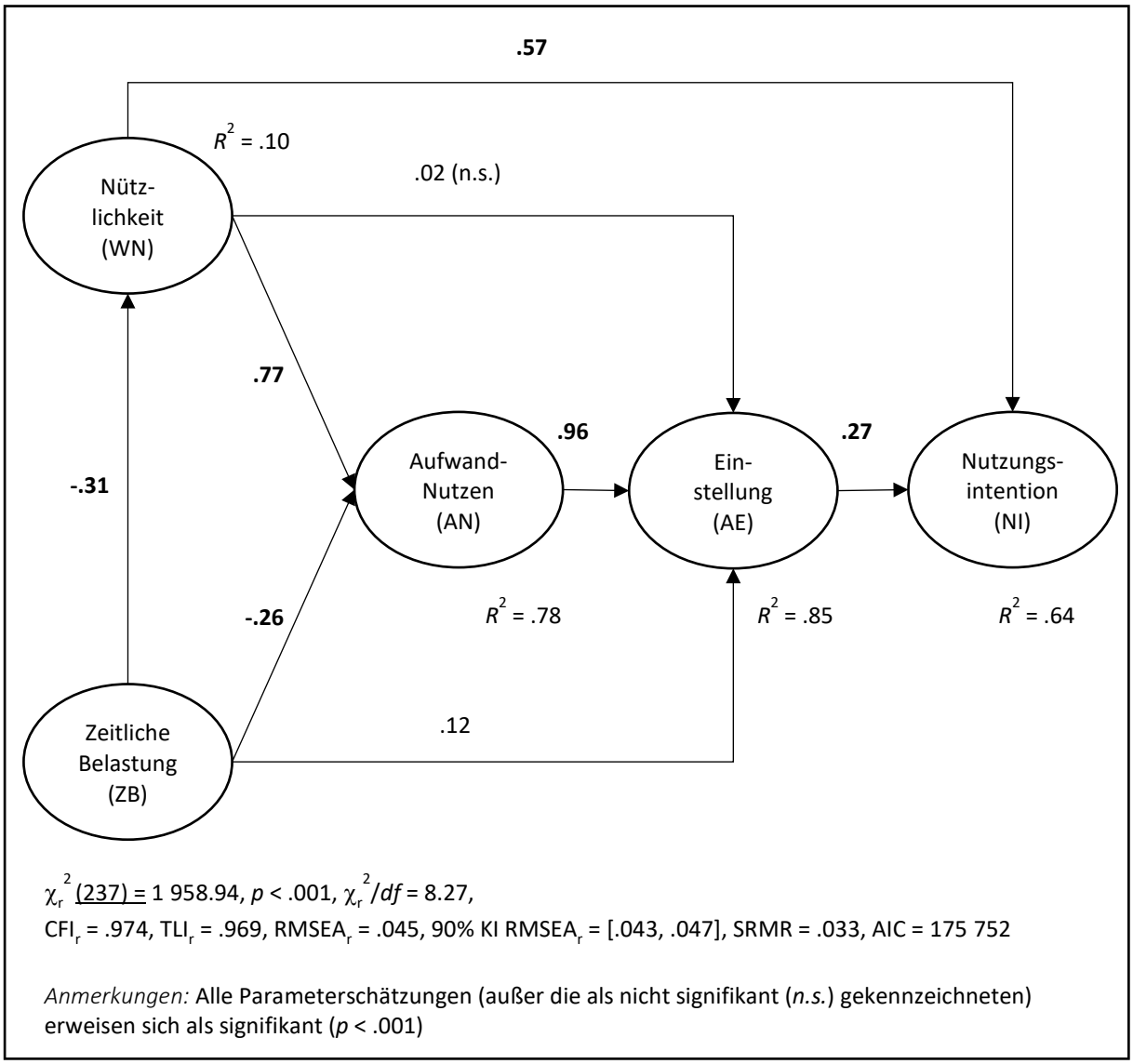


Abbildung 17: Modell $1_{dP(V3\text{-}18)}$ – Fitstatistiken und Parameterschätzungen (vollständig standardisierte Lösung).

Der Hauptunterschied in der Parameterschätzung zeigt sich bedingt durch den direkten Pfad der Nützlichkeit auf die Nutzungsintention. Dieser direkte positive Effekt β = .57 erweist sich als signifikant (p < .001) und führt zugleich zu einem verringerten direkten Einfluss der Einstellung auf die Nutzungsintention auf β = .27. Insgesamt nimmt der Anteil der durch das Modell erklärten Varianz der Nutzungsintention zu und steigt durch die Spezifizierung dieses direkten Effekts von 55 % auf 64 %. In der Gegenüberstellung der beiden Modelle passt Modell $1_{dP(V3\text{-}18)}$ besser zu den Daten und liefert deutliche Hinweise darauf, dass der wahrgenommenen Nützlichkeit eine wichtige Rolle bei der Nutzung von VERA-Daten zukommt.

Überprüfung der Hypothesen – direkte und indirekte Effekte

Einleitend zur Darstellung der direkten und indirekten Effekte und zur Beantwortung der in Kapitel 3 aufgestellten Hypothesen, sind diese zur Orientierung in Tabelle 35 noch einmal zusammengefasst dargestellt. In der zweiten Spalte ist jeweils die postulierte Wirkungsbeziehung zwischen zwei Konstrukten dargestellt, die weiteren Spalten differenzieren diese nach direktem (d), indirektem (i) und gesamtem (g) Effekt, jeweils für Modell $1_{(V3\text{-}18)}$ und Modell $1_{dP(V3\text{-}18)}$. Die Vorzeichen in den Zellen der Tabelle kennzeichnen jeweils das postulierte Vorzeichen des entsprechenden Effekts.

Die Stärke der analog zu diesen Hypothesen berechneten direkten und indirekten Effekte der Modelle $1_{(V3\text{-}18)}$ und $1_{dP(V3\text{-}18)}$ sind in Tabelle 36 abgebildet. Die Beantwortung der Hypothesen ist zusätzlich in Tabelle 37 übersichtlich dargestellt.

Wie bereits dargelegt, bestätigt sich die Vermutung eines direkten negativen Einflusses der zeitlichen Belastung auf die Wahrnehmung der Nützlichkeit ($\gamma_{M1(V3-18)}$ = -.32 bzw. $\gamma_{M1dP(V3-18)}$ = -.31) und liefert somit für beide Modelle Evidenz zur Annahme von Hypothese H1d/g. Mit der Wahrnehmung einer hohen zeitlichen Belastung sinkt demnach die Nutzenwahrnehmung des Instruments VERA.

Die Wirkung der zeitlichen Belastung auf die Abwägung von Aufwand und Nutzen erfolgt sowohl direkt (H2d, $\gamma_{M1(V3-18)}$ = -.25 bzw. $\gamma_{M1dP(V3-18)}$ = -.26) als auch indirekt, gemittelt über die wahrgenommene Nützlichkeit. Dieser indirekte Effekt beträgt -.25 bzw. -.24 (p < .001) und entspricht mit dem negativen Vorzeichen den Erwartungen eines ebenfalls negativen indirekten Effekts der zeitlichen Belastung auf das Aufwand-Nutzen-Konstrukt (H2i). Der gesamte Einfluss aus direktem und indirektem Effekt der zeitlichen Belastung auf die Einstellung liegt

sowohl bei Modell $1_{(V3\text{-}18)}$ als auch bei Modell $1_{dP(V3\text{-}18)}$ bei -.50 und ist auf dem 1 % Niveau signifikant (H2g). Hypothese 2 kann somit vollumfänglich anhand beider Modelle bestätigt werden, eine hohe empfundene zeitliche Belastung wirkt sowohl direkt als auch indirekt negativ auf die Aufwand-Nutzen-Abwägung von Lehrkräften.

Tabelle 35: Übersicht der aufgestellten Hypothesen

		Modell $1_{(V3\text{-}18)}$			Modell $1_{dP(V3\text{-}18)}$		
		Direkt (d)	Indirekt gesamt (i)	Gesamt (g)	Direkt (d)	Indirekt gesamt (i)	Gesamt (g)
H1	Zeitliche Belastung → Nützlichkeit	(−)		(−)	(−)		(−)
H2	Zeitliche Belastung → Aufwand-Nutzen	(−)	(−)	(−)	(−)	(−)	(−)
H3	Zeitliche Belastung → Einstellung	(−)	(−)	(−)	(−)	(−)	(−)
H4	Zeitliche Belastung → Nutzungsintention		(−)	(−)		(−)	(−)
H5	Nützlichkeit → Aufwand-Nutzen	(+)		(+)	(+)		(+)
H6	Nützlichkeit → Einstellung	(+)	(+)	(+)	(+)	(+)	(+)
H7	Nützlichkeit → Nutzungsintention		(+)	(+)	(+)	(+)	(+)
H8	Aufwand-Nutzen → Einstellung	(+)		(+)	(+)		(+)
H9	Aufwand-Nutzen → Nutzungsintention		(+)	(+)		(+)	(+)
H10	Einstellung → Nutzungsintention	(+)		(+)	(+)		(+)
H11	Effekt Nützlichkeit > Effekt zeitliche Belastung		(>)			(>)	

Anmerkungen: (-): negativer Effekt; (+): positiver Effekt; (>): stärkerer Effekt.

Wie im vorangegangenen Abschnitt beschrieben, zeigt sich entgegen den Erwartungen jeweils ein positiver direkter Effekt der zeitlichen Belastung auf die Einstellung, weshalb Hypothese H3d abgelehnt werden muss. Jedoch zeigt sich ein indirekter negativer Gesamteffekt (H3i). Die Summe der indirekten Effekte, mediiert über verschiedene Pfade zwischen Nützlichkeit und

Aufwand-Nutzen, erweist sich als signifikant ($p < .001$). Der erwartungsgemäß negative Effekt liegt in beiden Modellen bei -.48. Der Gesamteffekt, der durch den positiven direkten Effekt etwas abgeschwächt wird, ist aber insgesamt konform zur Hypothese H3g und liegt in beiden Modellen bei -.37 ($p < .001$). Die Hypothese H3 kann somit teilweise betätigt werden. Zwar findet sich kein direkter negativer Effekt einer hohen zeitlichen Belastung auf die Einstellung (H3d), dennoch übt eine hohe zeitliche Belastung durchaus einen negativen Einfluss auf die Einstellung der Lehrkräfte aus. Wie in den Hypothesen H3i und H3g postuliert, wirkt sie indirekt über Nützlichkeit und Aufwand-Nutzen-Abwägung nicht unwesentlich auf die Einstellung.

Die Vermutung eines negativen indirekten Gesamteffekts der zeitlichen Belastung auf die Nutzungsintention (H4i/g) erweist sich in beiden Modellen als zutreffend. Der indirekte Gesamteffekt liegt jeweils bei -.27 und ist auf dem 1 % Niveau signifikant. Die Hypothese H4i/g einer negativen Auswirkung hoher zeitlicher Belastung auf die Nutzungsintention der Ergebnisrückmeldungen kann somit bestätigt werden.

Der direkte Effekt der Nützlichkeit auf die Aufwand-Nutzen-Abwägung erweist sich in beiden Modellen konform mit der Hypothese H5d/g als signifikant positiv ($\beta = .77$, $p < .001$). Eine positive Nutzenwahrnehmung wirkt somit positiv auf die Aufwand-Nutzen-Bewertung durch die Lehrkräfte.

Der direkte Effekt der Nützlichkeit auf die Einstellung (H6d) hingegen wird zwar in Modell $1_{(V3\text{-}18)}$ auf dem 5 % Niveau signifikant, erweist sich aber als so marginal ($\beta = .06$), dass dessen Wirkung vernachlässigbar scheint. In Modell $1_{dP(V3\text{-}18)}$ wird der Effekt sogar insignifikant. Indes zeigt sich, wie in Hypothese 6i postuliert, ein starker signifikanter indirekter Effekt der wahrgenommenen Nützlichkeit, mediiert über die Aufwand-Nutzen-Abwägung von .72 in Modell $1_{(V3\text{-}18)}$ und .74 in Modell $1_{dP(V3\text{-}18)}$ ($p < .001$). Entsprechend fällt der Gesamteffekt mit .78 und .76 erwartungsgemäß positiv aus (H6g), was nahezu ausschließlich auf den beschriebenen indirekten Pfad zurückzuführen ist. Hypothese 6 kann somit nur teilweise bestätigt werden. Zwar gibt es insgesamt einen starken positiven Einfluss der wahrgenommenen Nützlichkeit der Vergleichsarbeiten auf die Einstellung der Lehrkräfte, jedoch ist dieser Effekt hauptsächlich durch den indirekten Effekt über den Einfluss der Nützlichkeit auf die Aufwand-Nutzen-Bewertung bestimmt. Die Gesamthypothese einer positiven Einstellung bei stärker ausgeprägter wahrgenommener Nützlichkeit (H6g) kann somit in beiden Modellen bestätigt werden, ebenso die Hypothese H6i.

Tabelle 36: Direkte und indirekte Effekte der Modelle $1_{(V3\text{-}18)}$ und $1_{dP(V3\text{-}18)}$ (vollständig standardisierte Lösung)

	Modell $1_{(V3\text{-}18)}$			Modell $1_{dP(V3\text{-}18)}$		
	Std.	*S. E.*	*z*-Wert	Std.	*S. E.*	*z*-Wert
Zeitliche Belastung → Nützlichkeit (H1)						
Direkt/Gesamteffekt (H1d/g)	-.317	.058	-9.602	-.309	.058	-9.374
Zeitliche Belastung → Aufwand-Nutzen (H2)						
Direkt (H2d)	-.250	.040	-10.891	-.256	.039	-11.216
Indirekt gesamt (H2i)	-.246	.043	-9.978	-.239	.042	-9.750
Gesamteffekt (H2g)	-.496	.070	-12.178	-.495	.070	-12.193
Zeitliche Belastung → Einstellung (H3)						
Direkt (H3d)	.115	.035	7.088	.115	.036	6.922
Indirekt gesamt (H3i)	-.480	.088	-11.967	-.481	.088	-11.998
Gesamteffekt (H3g)	-.365	.075	-10.640	-.366	.075	-10.719
Zeitliche Belastung → Nutzungsintention (H4)						
Indirekt gesamt/Gesamteffekt (H4i/g)	-.271	.049	-10.452	-.274	.052	-9.975
Nützlichkeit → Aufwand-Nutzen (H5)						
Direkt/Gesamteffekt (H5d/g)	.774	.016	46.713	.772	.016	47.305
Nützlichkeit → Einstellung (H6)						
Direkt (H6d)	.056[a]	.035	1.997	.020[b]	.035	0.697
Indirekt gesamt (H6i)	.722	.036	24.738	.740	.036	25.266
Gesamteffekt (H6g)	.778	.018	54.182	.759	.018	52.976
Nützlichkeit → Nutzungsintention (H7)						
Direkt (H7d)	-	-	-	.572	.027	22.515
Indirekt gesamt (H7i)	.578	.016	39.757	.202	.020	10.893
Gesamteffekt (H7g)				.775	.016	50.408
Aufwand-Nutzen → Einstellung (H8)						
Direkt/Gesamteffekt (H8d/g)	.932	.043	27.310	.959	.044	27.650
Aufwand-Nutzen → Nutzungsintention (H9)						
Indirekt gesamt/Gesamteffekt (H9i/g)	.693	.030	25.460	.255	.028	10.112
Einstellung → Nutzungsintention (H10)						
Direkt/Gesamteffekt (H10d/g)	.743	.012	52.370	.266	.021	10.958

Anmerkungen. $N = 4\,141$ (VERA3 2018); alle Parameterschätzungen erweisen sich als signifikant ($p < .001$), außer [a]$p = .046$ und [b]*n.s.*: nicht signifikant. Std.: standardisiert.

Gemäß Hypothese H7 steigt mit zunehmender Nutzenwahrnehmung die Intention zur Nutzung der Rückmeldungen für weitere Unterrichtsarbeit. Modell $1_{(V3\text{-}18)}$ liefert entsprechend der Modellspezifikation lediglich Evidenz für einen indirekten Effekt der Nützlichkeit auf die Nutzungsintention, mediiert durch Aufwand-Nutzen-Bewertung und Einstellung (H7i). Dieser Pfad wird auf dem 1 % Niveau signifikant und beträgt .58. In Modell $1_{dP(V3\text{-}18)}$ fällt dieser indirekte Effekt mit .20 deutlich geringer aus, bleibt jedoch signifikant. In Modell $1_{dP(V3\text{-}18)}$ zeigt sich hingegen analog zur Hypothese H7d ein relativ starker direkter Einfluss der Nützlichkeit auf die Nutzungsintention (β = .57, p < .001). Der Gesamteffekt aus direktem Effekt und indirekten Effekten liegt in diesem Modell bei .78 (H7g). Hypothese 7 kann somit für beide Modelle bestätigt werden. In Modell $1_{(V3\text{-}18)}$ wird der Effekt hier vollständig durch Aufwand-Nutzen und Einstellung mediiert (Annahme von H7i und H7g). Dieser Mediatoreffekt nimmt in Modell $1_{dP(V3\text{-}18)}$ zwar deutlich ab und wird teils durch einen relativ starken direkten Effekt abgelöst, dennoch fallen sowohl der indirekte Effekt als auch der direkte Pfad ebenso wie der Gesamteffekt signifikant positiv aus, weshalb alle Teilhypothesen auch für Modell $1_{dP(V3\text{-}18)}$ angenommen werden.

Hypothese H8d/g kann in beiden Modellen klar bestätigt werden, da ein sehr starker positiver Effekt der Aufwand-Nutzen-Abwägung auf die Einstellung der Lehrkräfte besteht ($\beta_{M1(V3-18)}$ = .93, $\beta_{M1dP(V3-18)}$ = .96, p < .001). Eine zunehmend günstige Abwägung von Aufwand und Nutzen führt demnach zu einer positiveren Einstellung. Jedoch muss dieser sehr starke Effekt, der sich bereits bei der latenten Korrelation der beiden Faktoren in der Faktorenanalyse mit allen Konstrukten angedeutet hat, wie bereits erwähnt noch einmal genauer betrachtet werden, um möglichen konzeptionellen Missspezifikationen zu begegnen (siehe Kapitel 5.2).

Die Ergebnisse der Modellschätzung stützen auch die Hypothese H9i/g eines indirekten positiven Effektes der Aufwand-Nutzen-Abwägung auf die Nutzungsintention. In beiden Modellen führt eine positivere Aufwand-Nutzen-Bewertung zu einer signifikant stärkeren Nutzungsintention, wobei dieser Effekt in Modell $1_{(V3\text{-}18)}$ mit .69 stärker ausfällt als in Modell $1_{dP(V3\text{-}18)}$ mit .26, wo ein Teil dieses indirekten Effekts durch die Spezifikation des direkten Pfades vom Konstrukt Nützlichkeit auf die Nutzungsintention kompensiert wird. Insgesamt lässt sich die Hypothese H9i/g jedoch in beiden Modellen bestätigen.

Entsprechend der Annahme von Hypothese H10d/g zeigt sich in beiden Modellen mit einer positiveren Einstellung eine größere Nutzungsintention. Auch dieser Effekt ist in

Modell $1_{(V3\text{-}18)}$, analog zu dem zuvor beschriebenen indirekten Effekt der Aufwand-Nutzen-Abwägung, stärker ausgeprägt ($\beta_{M1(V3-18)} = .74$, $p < .001$) als in Modell $1_{dP(V3\text{-}18)}$ ($\beta_{M1dP(V3-18)} = .27$, $p < .001$). Dennoch kann die Hypothese H10d bzw. g für beide Modelle angenommen werden, und die Ergebnisse sprechen dafür, dass eine positive Einstellung die Nutzungsintention der VERA-Rückmeldungen begünstigt.

Tabelle 37: Zusammenfassende Darstellung der überprüften Hypothesen

		Modell $1_{(V3\text{-}18)}$			Modell $1_{dP(V3\text{-}18)}$		
		Direkt (d)	Indirekt gesamt (i)	Gesamt (g)	Direkt (d)	Indirekt gesamt (i)	Gesamt (g)
H1	Zeitliche Belastung → Nützlichkeit	(−)		(−)	(−)		(−)
H2	Zeitliche Belastung → Aufwand-Nutzen	(−)	(−)	(−)	(−)	(−)	(−)
H3	Zeitliche Belastung → Einstellung	(−)	(−)	(−)	(−)	(−)	(−)
H4	Zeitliche Belastung → Nutzungsintention		(−)	(−)		(−)	(−)
H5	Nützlichkeit → Aufwand-Nutzen	(+)		(+)	(+)		(+)
H6	Nützlichkeit → Einstellung	(+) [a]	(+)	(+)	(+)	(+)	(+)
H7	Nützlichkeit → Nutzungsintention		(+)	(+)	(+)	(+)	(+)
H8	Aufwand-Nutzen → Einstellung	(+)		(+)	(+)		(+)
H9	Aufwand-Nutzen → Nutzungsintention		(+)	(+)		(+)	(+)
H10	Einstellung → Nutzungsintention	(+)		(+)	(+)		(+)
H11	Effekt Nützlichkeit > Effekt zeitliche Belastung		>			>	

Anmerkungen: (-): negativer Effekt; (+): positiver Effekt; (>): stärkerer Effekt.
Grün hinterlegte Zelle = Hypothese wird bestätigt; rot hinterlegte Zelle = Hypothese wird abgelehnt.
[a] Formal ist der Effekt zwar in Modell $1_{(V3\text{-}18)}$ auf dem 5 % Niveau signifikant, jedoch so marginal ($\beta = .06$), dass dessen Wirkung vernachlässigbar scheint; in Modell $1_{dP(V3\text{-}18)}$ wird der Effekt insignifikant; insgesamt wird die Hypothese zurückgewiesen.

Hypothese H11 postuliert in Anlehnung an die Erkenntnisse zu den Wirkungsbeziehungen des TAM (siehe bspw. Chismar & Wiley-Patton, 2003; King & He, 2006; Ma & Liu, 2004) einen im Vergleich zum Gesamteffekt der zeitlichen Belastung stärkeren Gesamteffekt der Nützlichkeit auf die Nutzungsintention. Diese Hypothese kann für beide Modelle bestätigt werden: Der Gesamteffekt der Nützlichkeit liegt in Modell $1_{(V3\text{-}18)}$ bei .58, in Modell $1_{dP(V3\text{-}18)}$ sogar bei .78, der Effekt der zeitlichen Belastung fällt dagegen mit jeweils .27 vergleichsweise gering aus. Insgesamt erweist sich die wahrgenommene Nützlichkeit als relevanter für die Nutzungsintention der VERA-Rückmeldungen als die empfundene zeitliche Belastung.

Zusammenfassend zeigt die detaillierte Analyse der Modellparameter mit Blick auf die Hypothesen, dass sich ein Großteil der im Voraus getroffenen Annahmen in beiden Modellen bestätigen. Die Hinzunahme des direkten Pfades von Nützlichkeit auf die Nutzungsintention beeinflusst hierbei insbesondere die Effekte der endogenen Variablen Nützlichkeit und Einstellung. Die Effekte der exogenen Variablen zeitliche Belastung bleiben nahezu unverändert, ebenso wie der direkte Effekt der Aufwand-Nutzen-Abwägung auf die Einstellung und die Effekte von Nützlichkeit auf Aufwand-Nutzen und Einstellung. Die deutliche Verbesserung der Modellgüte durch Einbezug des zusätzlichen direkten Pfades unterstreicht jedoch die Bedeutung der wahrgenommenen Nützlichkeit für die Intention zur Weiterarbeit mit den VERA-Ergebnissen (siehe hierzu die Diskussion der Ergebnisse in Kapitel 6).

Da die Parameterschätzungen noch einige Fragen aufwerfen, bspw. im Hinblick auf die starke Beziehung zwischen Aufwand-Nutzen-Abwägung und Einstellung, erfolgt im nächsten Schritt eine dahingehend datengeleitete Modellanpassung (siehe Kapitel 5.2).

5.2. Modellanpassung: Modell 2

Da der starke Zusammenhang zwischen der Aufwand-Nutzen-Abwägung und dem Konstrukt Einstellung die Vermutung nahelegt, dass die beiden Konstrukte für die Proband*innen nur schwer differenzierbar waren, wurde das ursprüngliche Modell angepasst und das Konstrukt Aufwand-Nutzen entfernt. Inhaltlich bleibt die Aufwand-Nutzen-Abwägung jedoch auch in diesem Modell enthalten, weil durch das Zusammenwirken der beiden Konstrukte zeitliche Belastung und Nutzenwahrnehmung eben diese Gegenüberstellung von Aufwand und Nutzen dennoch abgebildet wird. Lediglich auf eine Operationalisierung als eigenständiges Konstrukt wird in dem angepassten Modell verzichtet. Das angepasste Modell 2 entspricht somit im Hinblick

auf die spezifizierten Wirkungszusammenhänge stärker dem herkömmlichen TAM. Eine alternative Möglichkeit wäre die Entfernung des Konstrukts Einstellung, was jedoch wenig sinnvoll erscheint, da dadurch die affektive Komponente des Modells wegfallen würde und stattdessen die kognitive Aufwand-Nutzen-Abwägung zweifach, direkt und indirekt operationalisiert, beibehalten würde. Eine weitere Option wäre, wie bereits in Kapitel 5.1.2 (CFA mit allen Konstrukten) diskutiert, die Konstrukte Einstellung und Aufwand-Nutzen in einem gemeinsamen Faktor zu modellieren. Da dies jedoch, wie bereits dargelegt, zu einer Verschlechterung des Modellfits führt und zudem inhaltlich nicht plausibel ist, wurde auch von dieser Variante abgesehen.

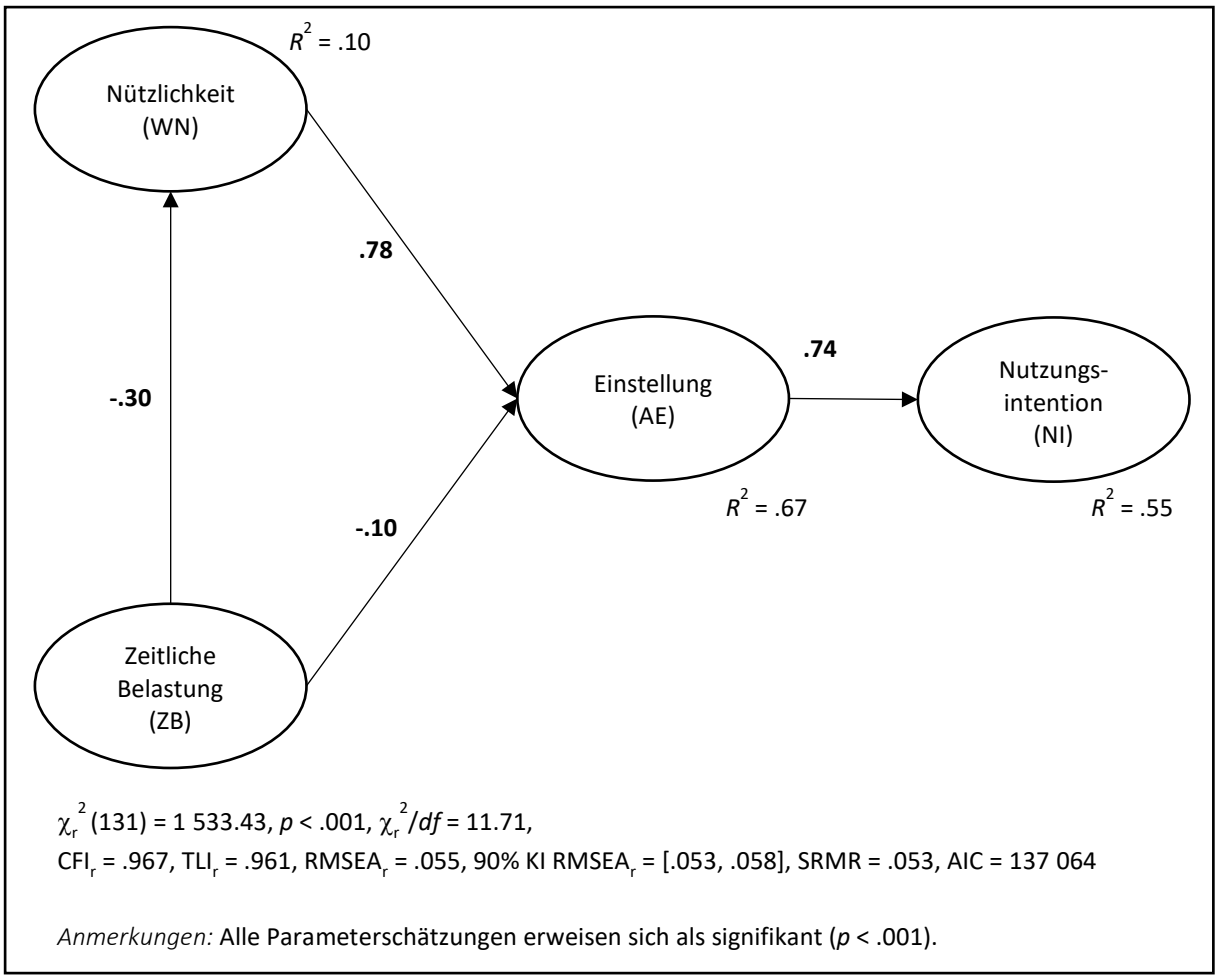


Abbildung 18: Modell $2_{(V3\text{-}18)}$ – Fitstatistiken und Parameterschätzungen (vollständig standardisierte Lösung).

Analog zu Modell $1_{(V3\text{-}18)}$ wurde in Modell $2_{(V3\text{-}18)}$ zunächst kein direkter Pfad zwischen wahrgenommener Nützlichkeit und Nutzungsintention definiert (siehe Abbildung 18). Die Fitstatistiken des Modells sprechen für eine insgesamt annehmbare Modellgüte. Die Indikatoren weisen alle einen akzeptablen Modellfit aus: CFI_r = .967, TLI_r = .961, $RMSEA_r$ = .055, SRMR = .053.

Insgesamt ist die Modellpassung vergleichbar mit der von Modell $1_{(V3\text{-}18)}$, $RMSEA_r$ und SRMR fallen im angepassten Modell etwas schlechter aus, CFI_r und TLI_r geringfügig besser.

Die Spezifikation der direkten Beziehung zwischen wahrgenommener Nützlichkeit und Nutzungsintention führt wiederum, wie in Modell $1_{dP(V3\text{-}18)}$, zu einer deutlichen Verbesserung der Modellgüte (siehe Modell $2_{dP(V3\text{-}18)}$, Abbildung 19). Alle Fitstatistiken sprechen durchweg für einen guten Modellfit: $CFI_r = .979$, $TLI_r = .975$, $RMSEA_r = .044$, SRMR = .034. Auch der AIC weist darauf hin, dass Modell $2_{dP(V3\text{-}18)}$ gegenüber Modell $2_{(V3\text{-}18)}$ zu bevorzugen ist und der zusätzlich zu schätzende Parameter den Modellfit verbessert ($AIC_{Modell2(V3\text{-}18)} = 137\,064$, $AIC_{Modell2dP(V3\text{-}18)} = 136\,465$) und das Modell auch im Vergleich zu Modell $1_{dP(V3\text{-}18)}$ ($AIC_{Modell1dP(V3\text{-}18)} = 175\,752$) zu bevorzugen wäre. Insgesamt kann die Modellgüte als gut betrachtet und das Modell beibehalten werden.

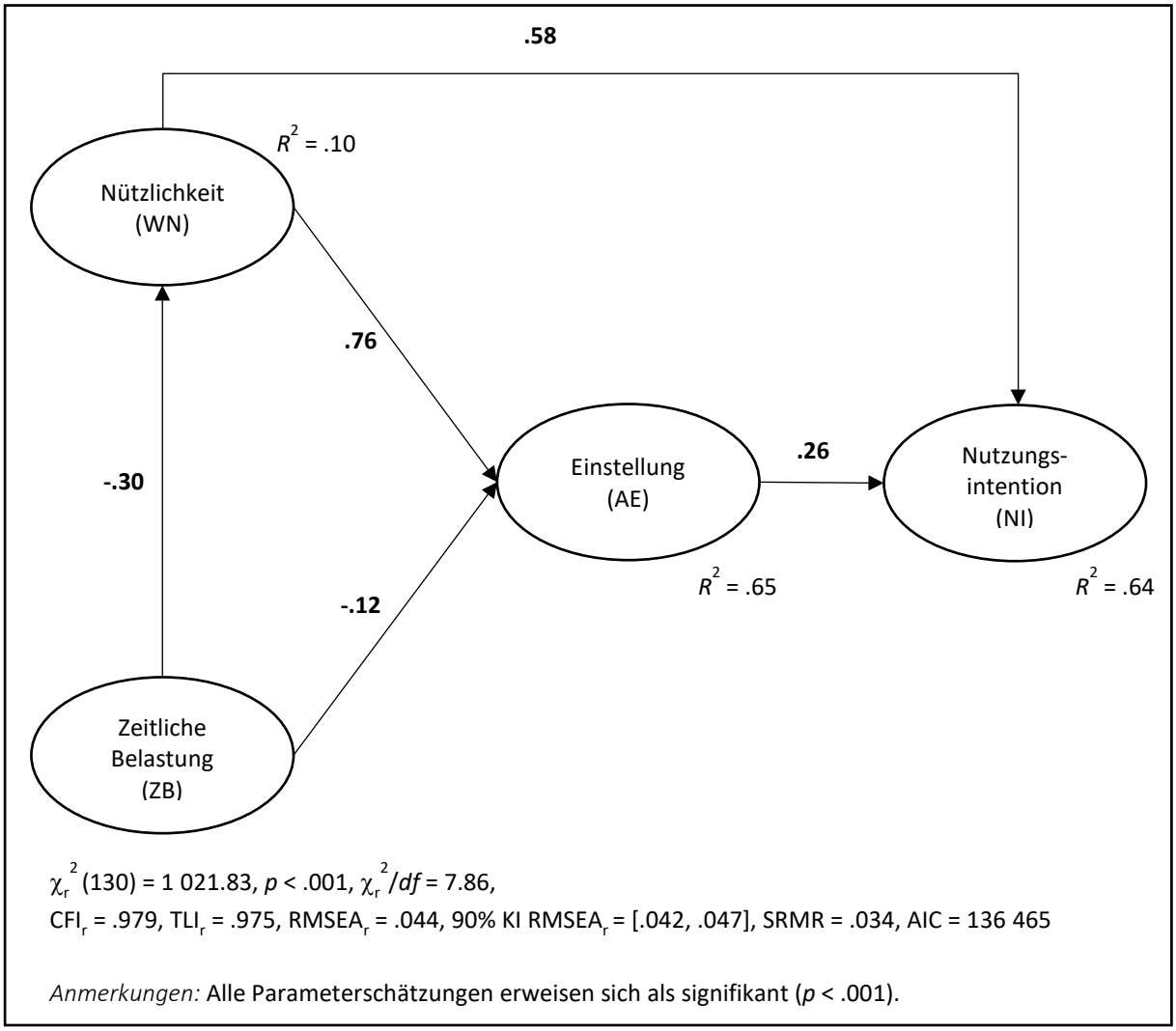


Abbildung 19: Modell $2_{dP(V3\text{-}18)}$ – Fitstatistiken und Parameterschätzungen (vollständig standardisierte Lösung).

Die Parameterschätzungen der jeweiligen Messmodelle von Modell $2_{(V3\text{-}18)}$ und Modell $2_{dP(V3\text{-}18)}$ sind in Anhang D, Tabelle 55 zu finden.

In beiden Modellen erweisen sich sämtliche Pfadkoeffizienten auf dem 1 % Niveau als signifikant und weisen das vermutete Vorzeichen auf. Tabelle 38 gibt dahingehend eine ausführliche Übersicht über die direkten und indirekten Effekte in Modell $2_{(V3\text{-}18)}$ und Modell $2_{dP(V3\text{-}18)}$. Der direkte Effekt der zeitlichen Belastung auf die wahrgenommene Nützlichkeit liegt in beiden Modellen bei -.30 und verweist auch im angepassten Modell auf eine abnehmende Nutzenwahrnehmung mit steigender zeitlicher Belastung. Die Stärke dieses Einflusses ist mit Modell $1_{(V3\text{-}18)}$ vergleichbar, ebenso wie die Varianzaufklärung. Der Einfluss der zeitlichen Belastung erklärt in allen Modellen 10 % der Varianz der wahrgenommenen Nützlichkeit (R^2 = .10). Ohne die zusätzlichen durch die Aufwand-Nutzen-Abwägung vermittelten indirekten Effekte wird der direkte Einfluss der zeitlichen Belastung auf die Einstellung in beiden Modellen wie erwartet signifikant negativ ($\gamma_{M2(V3-18)}$ = -.10 bzw. $\gamma_{M2dP(V3-18)}$ = -.12). Zusätzlich wirkt die zeitliche Belastung, mediiert über die wahrgenommene Nützlichkeit in beiden Modellen, indirekt auf die Einstellung der Lehrkräfte. Dieser indirekte Effekt liegt bei -.23 in Modell $2_{(V3\text{-}18)}$ und -.22 in Modell $2_{dP(V3\text{-}18)}$. Durch den Wegfall der weiteren indirekten Pfade über das Konstrukt Aufwand-Nutzen fällt der indirekte Gesamteffekt jeweils etwas geringer aus als in Modell 1, der Gesamteffekt aus direktem und indirektem Pfad ist aber mit Modell 1 vergleichbar und fällt in Modell $2_{(V3\text{-}18)}$ mit -.34 etwas kleiner aus als in Modell $1_{(V3\text{-}18)}$, entspricht jedoch in Modell $2_{dP(V3\text{-}18)}$ mit -.37 exakt dem Effekt in Modell $1_{dP(V3\text{-}18)}$. Der gesamte indirekte Effekt der zeitlichen Belastung liegt in beiden Modellen ($2_{(V3\text{-}18)}$ und $2_{dP(V3\text{-}18)}$) bei -.25 und ist somit nur geringfügig niedriger als im ursprünglichen Modell.

Im Gegensatz zu dem Ursprungsmodell, in dem nahezu der gesamte Effekt der Nützlichkeit auf die Einstellung durch das Aufwand-Nutzen-Konstrukt mediiert wurde, erweist sich die wahrgenommene Nützlichkeit in Modell 2 als signifikanter direkter Prädiktor der Einstellung. Die Höhe der direkten Effekte in den Modellen $2_{(V3\text{-}18)}$ und $2_{dP(V3\text{-}18)}$ entspricht mit $\beta_{M2(V3-18)}$ = .78 bzw. $\beta_{M2dP(V3-18)}$ = .77 hier in etwa der Höhe des (indirekten) Gesamteffekts in Modell $1_{(V3\text{-}18)}$ bzw. $1_{dP(V3\text{-}18)}$. Insgesamt erklärt das angepasste Modell 67 % (Modell $2_{(V3\text{-}18)}$) bzw. 65 % (Modell $2_{dP(V3\text{-}18)}$) der Varianz des latenten Konstrukts Einstellung. Die Varianzaufklärung der Einstellung ist somit rund 20 % geringer als im Ursprungsmodell, was vermutlich auf den Wegfall der starken Beziehung zu dem vorgelagerten Konstrukt Aufwand-Nutzen in jenem Modell zurückzuführen ist.

Tabelle 38: Direkte und indirekte Effekte der Modelle $2_{(V3\text{-}18)}$ und $2_{dP(V3\text{-}18)}$ (vollständig standardisierte Lösung)

	Modell $2_{(V3\text{-}18)}$			Modell $2_{dP(V3\text{-}18)}$		
	S. E.	*z*-Wert	Std.	*S. E.*	*z*-Wert	Std.
Zeitliche Belastung → Nützlichkeit						
Direkt/Gesamteffekt	.055	-9.353	-.296	.055	-9.127	-.288
Zeitliche Belastung → Einstellung						
Direkt	.037	-6.105	-.103	.038	-6.702	-.115
Indirekt gesamt	.053	-9.592	-.232	.051	-9.373	-.220
Gesamteffekt	.072	-10.163	-.335	.072	-10.203	-.366
Zeitliche Belastung → Nutzungsintention						
Indirekt gesamt/Gesamteffekt	.047	-9.977	-.248	.049	-9.657	-.254
Nützlichkeit → Einstellung						
Direkt/Gesamteffekt	.018	54.881	.784	.018	53.573	.765
Nützlichkeit → Nutzungsintention						
Direkt	-	-	-	.027	22.649	.575
Indirekt gesamt	.016	39.923	.582	.020	10.855	.202
Gesamteffekt				.016	50.342	.778
Einstellung → Nutzungsintention						
Direkt/Gesamteffekt	.012	51.928	.742	.021	10.930	.264

Anmerkungen. $N = 4\,141$ (VERA3 2018); Std.: standardisiert; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Die Beziehung zwischen Nützlichkeit und Nutzungsintention wird in Modell $2_{(V3\text{-}18)}$ analog zu Modell $1_{(V3\text{-}18)}$ nur durch indirekte Einflüsse bestimmt. Vermittelt durch das Einstellungskonstrukt beträgt der indirekte Effekt hier .58 und entspricht somit dem gesamten indirekten Effekt in Modell $1_{(V3\text{-}18)}$. Auch die Varianzaufklärung der Nutzungsintention ist in beiden Modellen identisch und liegt bei 55 % ($R^2 = .55$). Durch die Spezifikation des direkten Pfades kommt in Modell $2_{dP(V3\text{-}18)}$ entsprechend der direkte Effekt zwischen Nützlichkeit und Nutzungsintention hinzu, der vergleichbar mit dem Effekt in Modell $1_{dP(V3\text{-}18)}$ bei $\beta = .58$ liegt. Der indirekte Effekt verringert sich ebenso wie im Ursprungsmodell auf .20. Der Gesamteffekt nimmt entsprechend

gegenüber Modell $2_{(V3\text{-}18)}$ zu und liegt bei .78. Die gesamte Varianzaufklärung der Nutzungsintention beträgt in Modell $2_{dP(V3\text{-}18),}$ wie bereits in Modell $1_{dP(V3\text{-}18)}$, 64 % (R^2 = .64).

Auch der direkte Pfad zwischen Einstellung und Nutzungsintention verhält sich im angepassten Modell 2 ebenso wie im Ursprungsmodell 1. Liegt der direkte Effekt in Modell $2_{(V3\text{-}18)}$ noch bei β = .74, nimmt dieser direkte Einfluss mit der Spezifikation des zusätzlichen direkten Pfades zwischen Nützlichkeit und Nutzungsintention in Modell $2_{dP(V3\text{-}18)}$ deutlich ab (β = .26).

Hinsichtlich der Gesamtwirkung von zeitlicher Belastung und wahrgenommener Nützlichkeit auf die Nutzungsintention erweist sich auch in dem angepassten Modell die Nützlichkeit als der stärkere Prädiktor. Wie im Ursprungsmodell liegt der Gesamteffekt der wahrgenommenen Nützlichkeit bei .58 (Modell $2_{(V3\text{-}18)}$) bzw. .78 (Modell $2_{dP(V3\text{-}18)}$) und der Effekt der zeitlichen Belastung nur bei -.25.

Die gegenüber den ursprünglichen Modellen verbesserte Modellgüte der angepassten Modelle 2 und 2_{dP} lässt den Schluss zu, dass dieses Modell die Daten ebenso gut und teilweise sogar besser erklärt als das ursprünglich aufgestellte Modell. Dies gilt vor allem für Modell $2_{dP(V3\text{-}18)}$ mit dem direkten Pfad zwischen Nützlichkeit und Nutzungsintention, was erneut die direkte Relevanz der wahrgenommenen Nützlichkeit für die Intention zur Weiterarbeit, auch unabhängig von der Einstellung der Lehrkräfte, unterstreicht. Zudem zeigen sich im Gegensatz zum Ursprungsmodell keine unplausiblen Effektschätzungen, da der Problematik einer möglichen Nicht-Differenzierbarkeit der Konstrukte Einstellung und Aufwand-Nutzen durch Entfernen des Konstrukts Aufwand-Nutzen begegnet wurde. Darüber hinaus führt Modell 2 zu analogen Effekten bei einer insgesamt besseren Modellpassung und einer effizienteren Modellschätzung.

Zusammenfassend lassen die gute Modellpassung, insbesondere von Modell $2_{dP(V3\text{-}18)}$, sowie die Evaluation der Modellparameter den Schluss zu, dass die in diesem Modell gewählte indirekte Konzeptualisierung der Aufwand-Nutzen-Abwägung durch die Lehrkräfte eine adäquate Modellierung des Phänomens abzubilden scheint. Eine explizite Operationalisierung wie im Ursprungsmodell ist demnach nicht unbedingt notwendig, wenn es das Ziel ist, die Nutzungsintention bzw. Intention zur Weiterarbeit mit den VERA-Rückmeldungen zu erklären. Daher wird das angepasste Modell 2 beibehalten und im nächsten Schritt mit einem weiteren unabhängigen Datensatz validiert (siehe Kapitel 5.3).

5.3. Modellvalidierung (VERA3 2019)

Da im Jahr 2019 eine mit der Vorjahresbefragung vergleichbare Evaluationsstudie mit VERA3-Lehrkräften durchgeführt wurde, konnte das angepasste Modell 2 mit dieser Stichprobe validiert werden. Die Ergebnisse dieser Modellvalidierung werden in diesem Kapitel berichtet. Der Fokus liegt hier vor allem auf der Herausarbeitung von Abweichungen der Modellschätzung mit der ursprünglichen Stichprobe, eine umfassende Darstellung der Ergebnisse findet sich in Anhang E.

Analog zum Vorgehen der Modelltestung mit dem ursprünglichen Datensatz wurden auch für diese Stichprobe im ersten Schritt die deskriptiven Statistiken der Indikatorvariablen betrachtet, u. a. mit dem Ziel, Abweichungen von einer multivariaten Normalverteilung aufzudecken. Wie bei der Stichprobe des Vorjahres kann auch hier nicht von einer multivariaten Normalverteilung ausgegangen werden, da sowohl Schiefe und Kurtosis als auch der für alle Items signifikante Shapiro-Wilk-Test auf eine Abweichung von der Normalverteilung hinweisen. Ausführlich ist die deskriptive Auswertung in Tabelle 56 (Anhang E) abgebildet. Basierend auf dieser Erkenntnis wurden die Messmodelle der für Modell 2 relevanten latenten Konstrukte entsprechend dem bisherigen Vorgehen zunächst mittels MLR-Schätzmethode jeweils separat geschätzt und dann in einem gemeinsamen Messmodell getestet, bevor das Strukturmodell (Modell $2_{(V3\text{-}19)}$ und Modell $2_{dP(V3\text{-}19)}$) untersucht wurde. Die konfirmatorischen Faktorenanalysen der einzelnen Konstrukte einschließlich weiterer Reliabilitätsanalysen sind in Anhang E (siehe Tabelle 57 bis Tabelle 67) ausführlich dargestellt. Die Durchführung der CFAs mit der VERA3-Stichprobe 2019 führt zu sehr ähnlichen Ergebnissen wie mit der Ursprungsstichprobe. Alle Messmodelle lassen sich mit dem neuen Datensatz abbilden, wobei sich auch die bereits offengelegten Schwächen der Messmodelle, wie die nicht optimale Passung des Konstrukts zeitliche Belastung, ebenfalls in dieser neuen Stichprobe zeigen. Insgesamt ist die Modellpassung der Messmodelle in der VERA3-Stichprobe 2019 etwas besser als in der ursprünglichen Stichprobe (siehe Tabelle 39).

Analog zu dem in Kapitel 5.1.2 gewählten Vorgehen wurden nach der Durchführung einzelner Faktorenanalysen alle Konstrukte in einem Messmodell geschätzt. Tabelle 39 enthält die Fitstatistiken dieser Modellschätzung im Vergleich zu den Fitindikatoren des Modells mit den Daten der VERA3-Stichprobe 2018. Auch mit dieser Stichprobe weisen alle Indikatoren auf einen guten Modellfit hin: $CFI_r = .983$, $TLI_r = .980$, $RMSEA_r = .040$, $SRMR = .032$. Die geschätzten

Modellparameter (Faktorladung, Standardfehler, z-Werte und Varianzaufklärung) sind in Tabelle 68 (siehe Anhang E) dargestellt.

Tabelle 39: Fitstatistiken der konfirmatorischen Faktorenanalyse mit allen Konstrukten aus Modell 2 im Vergleich VERA3 2018 und VERA3 2019

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
$2_{CFA(V3-18)}$	1 012.39 (.000)	129	7.85	**.979**	**.975**	**.044**	[.041, .047]	**.033**
$2_{CFA(V3-19)}$	627.24 (.000)	129	4.86	**.983**	**.980**	**.040**	[.037, .043]	**.032**

Anmerkungen. $N = 4\,141$ (VERA3 2018), $N = 2751$ (VERA3 2019); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Mit Blick auf die latenten Korrelationen zwischen den Faktoren (siehe Tabelle 40) fällt auf, dass die Konstrukte Nutzungsintention und wahrgenommene Nützlichkeit mit .87 ($p < .001$) sehr hoch miteinander korrelieren. Da der Wert jedoch noch unter der kritischen Grenze von $r = .90$ (Kline, 2011) liegt, werden beide Konstrukte unverändert beibehalten. Im Vergleich zu der Modelltestung mit dem ursprünglichen Datensatz ist diese Korrelation .08 stärker, was sich später auch in der Parameterschätzung des Strukturmodells niederschlägt. Die Korrelation zwischen Einstellung und Nützlichkeit ist mit $r = .80$ in beiden Stichproben identisch, die Korrelationen zwischen dem Konstrukt zeitliche Belastung und den anderen Konstrukten fällt dagegen durchweg etwas geringer aus als im ersten Datensatz, ebenso die Korrelation zwischen Einstellung und Nutzungsintention.

Tabelle 40: Korrelationen der latenten Faktoren (Modell $2_{CFA(V3-19)}$) (vollständig standardisierte Lösung) sowie manifeste Skalenmittelwerte und Standardabweichungen

		1.	2.	3.	4.	*M*	*SD*	*n*
1.	Nutzungsintention [a]					2.46	0.71	2 548
2.	Einstellung [a]	.70				2.38	0.86	2 631
3.	Zeitliche Belastung [b]	-.20	-.32			3.10	0.65	2 681
4.	Nützlichkeit [a]	.87	.80	-.24		2.39	0.73	2 571

Anmerkungen. $N = 2\,751$ (VERA3 2019); [a] 4-stufige positiv gepolte Antwortskala; [b] 5-stufige negativ gepolte Antwortskala.

Die manifesten Mittelwerte weisen weitestgehend vergleichbare Tendenzen zum ersten Datensatz auf: Der Mittelwert des Konstrukts Einstellung ist mit $M = 2.38$ ($SD = 0.86$) in beiden Stichproben identisch, während die zeitliche Belastung mit $M = 3.10$ ($SD = 0.65$) geringfügig größer wahrgenommen wird. Der Mittelwert der Nutzungsintention liegt mit $M = 2.46$ ($SD = 0.71$) .10 über dem der VERA3-Stichprobe 2018, der Mittelwert der Nützlichkeit fällt dagegen mit $M = 2.39$ ($SD = 0.73$) entsprechend .10 geringer aus.

Insgesamt sprechen die Ergebnisse der Modellschätzung für eine gute Modellpassung und die Gültigkeit des aufgestellten Messmodells unabhängig von der getesteten Stichprobe.

Auf die Untersuchung des Messmodells folgte die Schätzung des angepassten Strukturmodells (Modell $2_{(V3\text{-}19)}$ und $2_{dP(V3\text{-}19)}$). Die Fitstatistiken von Modell $2_{(V3\text{-}19)}$, zunächst ohne direkten Pfad zwischen Nützlichkeit und Nutzungsintention, und Modell $2_{dP(V3\text{-}19)}$, inklusive spezifiziertem direktem Pfad, sind in Tabelle 41 im Vergleich zur Schätzung mit den VERA3-Daten des Jahres 2018 nachzulesen. In Anhang E kann außerdem die grafische Darstellung der beiden Modelle nachgeschlagen werden (siehe Abbildung 35 und Abbildung 36), ebenso wie die Parameterschätzung der entsprechenden Messmodelle (siehe Tabelle 69). Vergleichbar mit der Analyse der VERA3-Daten aus 2018 zeigt sich durch die Spezifikation des direkten Pfades eine deutliche Verbesserung des Modellfits. Ist die Modellgüte von Modell $2_{(V3\text{-}19)}$ gerade noch akzeptabel, weist Modell $2_{dP(V3\text{-}19)}$ hinsichtlich aller Gütekriterien eine gute Modellpassung auf.

Tabelle 41: Fitstatistiken der Strukturmodelle $2_{(V3\text{-}19)}$ und $2_{dP(V3\text{-}19)}$ der Validierungsstichprobe VERA3 2019 im Vergleich zu den Strukturmodellen $2_{(V3\text{-}18)}$ und $2_{dP(V3\text{-}18)}$ der Stichprobe VERA3 2018

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	AIC
$2_{(V3\text{-}19)}$	1 411.03 (.000)	131	10.77	*.956*	.949	*.064*	[.061, .067]	*.068*	92 578
$2_{(V3\text{-}18)}$	1 533.43 (.000)	131	11.71	*.967*	*.961*	*.055*	[.053, .058]	*.053*	137 064
$2_{dP(V3\text{-}19)}$	628.12 (.000)	130	4.83	**.983**	**.980**	**.040**	[.037, .043]	**.032**	**91 672**
$2_{dP(V3\text{-}18)}$	1021.83 (.000)	130	7.86	**.979**	**.975**	**.044**	[.042, .047]	**.034**	**136 465**

Anmerkungen. $N = 2\,751$ (VERA3 2019); $N = 4\,141$ (VERA3 2018); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Verglichen mit der VERA3-Stichprobe 2018 fällt die Modellgüte von Modell $2_{(V3\text{-}19)}$ hier etwas schlechter aus, die Passung von Modell $2_{dP(V3\text{-}19)}$ hingegen besser als bei der ursprünglichen Stichprobe. Eindeutig ist jedoch der Trend des verbesserten Modellfits durch das Definieren des direkten Pfades zwischen Nützlichkeit und Nutzungsintention.

Tabelle 42: Direkte und indirekte Effekte der Modelle $2_{(V3\text{-}19)}$ und $2_{dP(V3\text{-}19)}$ (VEAR3 2019) (vollständig standardisierte Lösung)

	Modell $2_{(V3\text{-}19)}$			Modell $2_{dP(V3\text{-}19)}$		
	S. E.	*z*-Wert	Std.	*S. E.*	*z*-Wert	Std.
Zeitliche Belastung → Nützlichkeit						
Direkt/Gesamteffekt	.052	-7.161	-.240	.053	-7.085	-.237
Zeitliche Belastung → Einstellung						
Direkt	.039	-6.807	-.131	.039	-7.241	-.139
Indirekt gesamt	.053	-7.287	-.189	.051	-7.203	-.182
Gesamt	.069	-9.388	-.319	.069	-9.483	-.321
Zeitliche Belastung → Nutzungsintention						
Indirekt gesamt/Gesamteffekt	.042	-9.232	-.234	.049	-7.179	-.209
Nützlichkeit → Einstellung						
Direkt/Gesamteffekt	.022	46.736	.786	.021	46.165	.767
Nützlichkeit → Nutzungsintention						
Direkt	-	-	-	.031	28.426	.846
Indirekt gesamt	.019	32.737	.575	.023	0.972	*.021*[a]
Gesamt				.017	52.247	.867
Einstellung → Nutzungsintention						
Direkt/Gesamteffekt	.015	39.315	.731	.023	0.969	*.028*[a]

Anmerkungen. $N = 2\,751$ (VERA3 2019); Std.: standardisiert; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$); außer [a] *n.s.*: nicht signifikant.

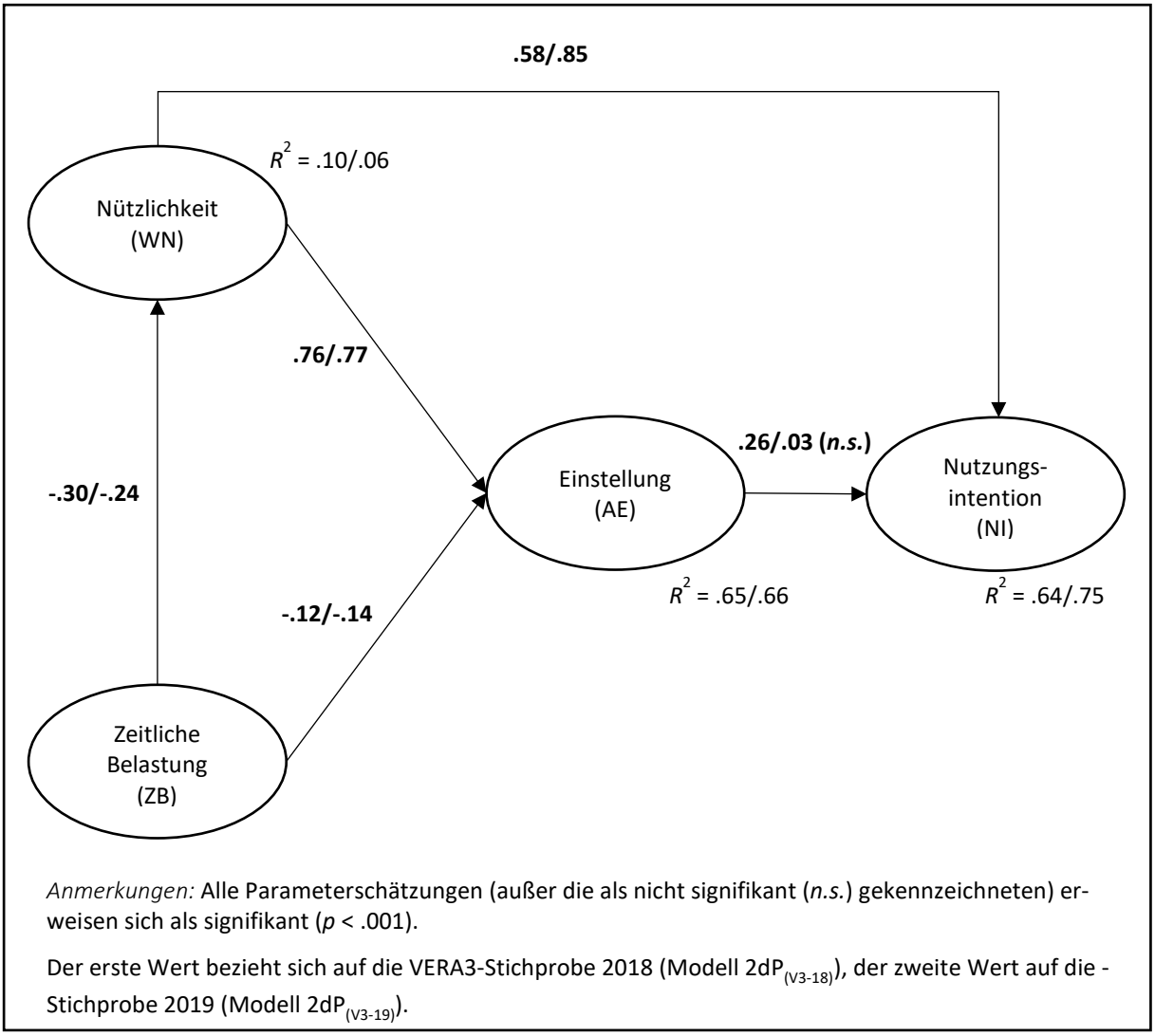
Auch hinsichtlich der Parameterschätzungen (siehe auch Anhang E Tabelle 69 und Tabelle 42) zeigen sich zwischen den Jahren vergleichbare Tendenzen in den Ausprägungen der Pfadkoeffizienten und der Varianzaufklärung der latenten Konstrukte. Dies gilt insbesondere für

Modell $2_{(V3\text{-}19)}$, für welches sich neben der Varianzaufklärung auch direkte und indirekte Effekte zwischen den Stichproben nur unwesentlich unterscheiden. Die größte Abweichung liegt hier mit einem Betrag von .06 beim Pfadkoeffizienten des direkten Effekts der zeitlichen Belastung auf die wahrgenommene Nützlichkeit. Modell $2_{dP(V3\text{-}19)}$ offenbart ebenfalls gleichartige Tendenzen hinsichtlich der Entwicklung der Stärke der Pfadkoeffizienten bei Hinzunahme des direkten Pfades.

Wie sich jedoch bereits durch die starke latente Korrelation zwischen Nützlichkeit und Nutzungsintention angedeutet hat, zeigt sich ein im Vergleich zur ursprünglichen Stichprobe (VERA3 2018) deutlich stärkerer Einfluss der Nützlichkeit auf die Nutzungsintention (β = .85, p < .001). Durch den direkten Pfad steigt die Varianzaufklärung der Nutzungsintention in Modell $2_{dP(V3\text{-}19)}$ auf 75 % und ist somit 11 % höher als 2018. Der direkte positive Effekt der Einstellung auf die Nutzungsintention wird gleichzeitig insignifikant (β = .03, n.s.). Dadurch verringert sich auch der indirekte Effekt der zeitlichen Belastung auf die Nutzungsintention im Vergleich zur anderen Stichprobe. Der indirekte Effekt der Nützlichkeit auf die Nutzungsintention wird nicht nur wie in der Stichprobe aus 2018 deutlich kleiner, sondern auch insignifikant (siehe Abbildung 20 für die vergleichende Darstellung von Modell 2_{dP} zwischen den Jahren 2018 und 2019).

Zusammenfassend kann festgehalten werden, dass die Validierung des angepassten Modells 2 zu der Erkenntnis führt, dass dieses Modell auch die Validierungsstichprobe angemessen widerspiegelt, da die Gütekriterien insgesamt einen akzeptablen (Modell $2_{(V3\text{-}19)}$) bzw. guten (Modell $2_{dP(V3\text{-}19)}$) Modellfit anzeigen. Auch die Parameterschätzungen sind zwischen den Jahren vergleichbar, sowohl auf Mess- als auch auf Strukturebene. Vor allem in Modell $2_{(V3\text{-}19)}$ unterscheiden sich die Effektschätzungen zwischen den Jahren nur marginal.

In Modell $2_{dP(V3\text{-}19)}$ ist die Kausalstruktur des Pfaddiagramms im Bereich der exogenen Variablen zeitliche Belastung und des direkten Effektes der Nützlichkeit auf die Einstellung ebenfalls zwischen den Jahren weitestgehend konstant. Auch mit den Daten aus 2019 verdeutlicht Modell $2_{dP(V3\text{-}19)}$ die Relevanz der direkten Beziehung zwischen Nützlichkeit und Nutzungsintention, wobei dieser Effekt hier noch stärker ausgeprägt ist als in der Ursprungsstichprobe. Die Erkenntnis bleibt die gleiche: Die Nutzenwahrnehmung von Vergleichsarbeiten erweist sich als wichtigster Prädiktor einer geplanten Weiterarbeit mit den durch die Testung gewonnenen Erkenntnissen.



Anmerkungen: Alle Parameterschätzungen (außer die als nicht signifikant (*n.s.*) gekennzeichneten) erweisen sich als signifikant ($p < .001$).

Der erste Wert bezieht sich auf die VERA3-Stichprobe 2018 (Modell $2dP_{(V3\text{-}18)}$), der zweite Wert auf die -Stichprobe 2019 (Modell $2dP_{(V3\text{-}19)}$).

Abbildung 20: Modell $2_{dP(V3\text{-}18)}$ – Parameterschätzungen der Modelle $2_{dP(V3\text{-}18)}$ und $2_{dP(V3\text{-}19)}$ (vollständig standardisierte Lösung).

Nachdem sich das angepasste Modell 2 auch in der Validierung mit einer unabhängigen Stichprobe bewährt hat, wird das Ursprungsmodell 1 für die weiteren Untersuchungen endgültig verworfen. Die folgenden Analysen zur Prüfung von Gruppenunterschieden zwischen VERA3 und VERA8 (siehe Kapitel 5.4) sowie die Diskussion der Ergebnisse in Kapitel 6 erfolgen auf Basis des angepassten Modells 2.

5.4. Latente Gruppenanalyse

Nachdem die Gültigkeit des angepassten Strukturmodells (Modell 2) in Kapitel 5.3 auch über mehrere Jahre hinweg für eine weitere unabhängige Stichprobe nachgewiesen werden konnte, beschäftigt sich das folgende Unterkapitel mit der ebenfalls in Kapitel 3 aufgeworfenen Frage

nach Unterschieden in der Wahrnehmung und Bewertung von Vergleichsarbeiten durch Lehrkräfte verschiedener Schulformen. Hierfür wurde die Passung des ermittelten Strukturmodells mit Hilfe einer Stichprobe aus VERA8-Lehrkräften überprüft. Mit dem Ziel, die Modellpassung und Parameterschätzungen zwischen VERA3- und VERA8-Lehrkräften zu analysieren und mögliche Unterschiede in der Bewertung der Vergleichsarbeiten zwischen Lehrkräften in der Grundschule und Lehrkräften der Sekundarstufe I zu untersuchen, wurde zu diesem Zweck ein latentes (Zwei-)Gruppenmodell aufgestellt. Die Analyse folgt den in Kapitel 4.3.4 beschriebenen Prüfschritten: Bevor etwaige Unterschiede des Strukturmodells und der latenten Mittelwerte untersucht werden konnten, mussten demnach die Messmodelle auf Messinvarianz inspiziert werden. Hierbei wurden erneut zunächst die Messmodelle der einzelnen Konstrukte geschätzt und dann im nächsten Schritt für alle Konstrukte ein gemeinsames Messmodell aufgestellt. Waren die entsprechenden Invarianzbedingungen erfüllt, wurden auf Basis eines gemeinsamen Messmodells Unterschiede in den latenten Mittelwerten betrachtet und in einem darauf basierenden Strukturmodell die Gleichheit der Pfadkoeffizienten untersucht. Datengrundlage dieser Untersuchung bilden eine Teilstichprobe der VERA3-Lehrkräftebefragung 2018 und die Evaluationsbefragung mit VERA8-Lehrkräften im Nachgang der VERA-Testung desselben Jahres (siehe auch Kapitel 4.2.5). Vor der latenten Gruppenanalyse wurden erneut die im folgenden Kapitel beschriebenen Item- und Skalenkennwerte im Gruppenvergleich inspiziert.

5.4.1. Itemkennwerte und Ergebnisse der Reliabilitätsanalyse

Tabelle 43 enthält die Mittelwerte der Indikatorvariablen einschließlich Standardabweichung der betrachteten Gruppen sowie Cronbachs Alpha und die durchschnittlich erfasste Varianz der einzelnen Konstrukte. Bei den Konstrukten Nutzungsintention und Einstellung fällt die durchschnittliche Bewertung aller Indikatorvariablen in der VERA8-Gruppe niedriger aus als in der VERA3-Gruppe, die Mittelwerte aller Items liegen unterhalb des theoretischen Mittelwertes von 2.50. Nur Item NI5 fällt mit einer eher neutralen Bewertung bei einem Mittelwert von 2.49 (SD = 0.91) aus dieser negativen Einschätzung heraus. Ein vergleichbares Bild zeigt sich beim Konstrukt Nützlichkeit, auch dort ist die Einschätzung der Items in der VERA8-Gruppe durchweg negativer als in der VERA3-Stichprobe. Die Streuung fällt dagegen in der VERA8-Gruppe bei nahezu allen Items etwas höher aus als in der VERA3-Gruppe, bei den Items NI3 und NI4 ist die Standardabweichung zumindest identisch.

Bezüglich der Items des Konstrukts zeitliche Belastung sind die Bewertungstendenzen zwischen den Gruppen weniger eindeutig. Während VERA3-Lehrkräfte einen höheren Vorbereitungsaufwand empfinden (Item ZB1) (M = 2.71, SD = 0.90) als VERA8-Lehrkräfte (M = 2.44, SD = 1.04), fällt die Bewertung des empfundenen Arbeitsaufwandes der Auswertung (Item ZB3) umgekehrt aus (VERA3: M = 3.40, SD = 0.78; VERA8: M = 3.88, SD = 0.88), was eine höhere empfundene Auswertungsbelastung in der Sekundarstufe widerspiegelt. Die Einschätzungen hinsichtlich des Durchführungsaufwandes (Item ZB2) unterscheiden sich nur geringfügig zwischen den Gruppen (VERA3: M = 3.04, SD = 0.68; VERA8: M = 3.02, SD = 0.85).

Tabelle 43: Mittelwerte und Standardabweichungen sowie interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Indikatorvariablen im Vergleich VERA3 und VERA8

		VERA3 2018				VERA8 2018			
Konstrukt	Item	*M*	*SD*	α	DEV	*M*	*SD*	α	DEV
Nutzungsintention Min = 1, Max = 4 M_T = 2.50	NI1	2.33	0.84	.89	.62	2.16	0.86	.89	.62
	NI2	2.37	0.82			2.21	0.85		
	NI3	2.13	0.78			2.01	0.78		
	NI4	2.35	0.85			2.14	0.85		
	NI5	2.68	0.85			2.49	0.91		
Einstellung Min = 1, Max = 4 M_T = 2.50	AE1	2.40	0.90	.94	.77	2.29	0.96	.94	.76
	AE2	2.47	0.82			2.39	0.85		
	AE3	2.44	0.88			2.37	0.93		
	AE4	2.39	0.89			2.25	0.93		
	AE5	2.38	0.93			2.39	0.98		
Zeitliche Belastung Min = 1, Max = 5 (inv. Polung) M_T = 3.00	ZB1	2.71	0.90	.55	.30	2.44	1.04	.56	.31
	ZB2	3.04	0.68			3.02	0.85		
	ZB3	3.40	0.78			3.88	0.88		
Wahrgenommene Nützlichkeit Min = 1, Max = 4 M_T = 2.50	WN1	2.39	0.84	.89	.62	2.18	0.89	.90	.63
	WN2	2.36	0.88			2.33	0.93		
	WN3	2.43	0.91			2.27	0.93		
	WN4	2.49	0.89			2.39	0.93		
	WN5	2.39	0.84			2.44	0.94		

Anmerkungen. VERA8 2018: N = 782; VERA3 2018-Teilstichprobe: N = 1 918.

Cronbachs Alpha liegt in beiden Gruppen für alle Konstrukte über dem Grenzwert von .50, ebenso die durchschnittlich erfasste Varianz, abgesehen vom Konstrukt zeitliche Belastung. Die Werte von Cronbachs Alpha und der durchschnittlich erfassten Varianz je Konstrukt sind zwischen den Gruppen nahezu identisch und variieren maximal um .01.

Dieses Ergebnis liefert erste Hinweise darauf, dass die Indikatorvariablen in beiden Gruppen, von den bereits beschriebenen Mängeln des Konstrukts zeitliche Belastung abgesehen, geeignet sind, die untersuchten Konstrukte angemessen zu erfassen. Diese Untersuchung wird in Kapitel 5.4.2 mit der Invarianzprüfung fortgesetzt.

Eine detailliertere Darstellung der Itemkennwerte, einschließlich Verteilungsstatistiken und dem Anteil fehlender Werte je Item sowie die manifesten Korrelationen der Indikatorvariablen für die VERA8-Stichprobe 2018, ist in Anhang F beigefügt. Auffällig ist, dass der Anteil der fehlenden Werte bei den Items der Konstrukte Einstellung, zeitliche Belastung und Nutzungsintention in etwa denen in den beiden VERA3-Stichproben entsprechen. Der Anteil fehlender Werte für das Konstrukt Nützlichkeit liegt hingegen in der VERA8-Stichprobe mit 6.4 % (Item WN5) bis 9.1 % (Item WN1) deutlich höher als in den VERA3-Stichproben. Dort liegt ihr Anteil bei maximal 4.9 % (VERA3 2018) bzw. 4.4 % (VERA3 2019). Möglicherweise hatten die Lehrkräfte der Sekundarstufe mehr Probleme diese spezifischen Fragen nach der Nützlichkeit der VERA-Rückmeldungen zu beantworten, ohne diese aufgrund des Erhebungszeitpunktes zuvor gesehen zu haben, da bei VERA8 die Wahrscheinlichkeit einer Vorerfahrung mit Vergleichsarbeiten etwas geringer ist als bei VERA3. Grundschullehrkräfte sehen sich tendenziell häufiger mit der Durchführung der Vergleichsarbeiten konfrontiert, weil die Wahrscheinlichkeit, dass sie eine dritte Klasse unterrichten und somit an VERA3 teilnehmen, höher ist als die einer Sekundarschullehrkraft, eine achte Klasse zu unterrichten. Hinzu kommt, dass die verpflichtenden Testfächer in VERA8 zwischen den Jahren wechseln, was die Wahrscheinlichkeit für die Teilnahme an VERA8 für Sekundarschullehrkräfte weiter senkt. Dies könnte den größeren Anteil fehlender Werte in der VERA8-Stichprobe erklären.

5.4.2. Messinvarianzprüfung

Ziel der Messinvarianzprüfung war es, die Gültigkeit der in Kapitel 5.1.2 ermittelten Messmodelle für eine Stichprobe aus Sekundarschullehrkräften (VERA8) zu überprüfen. Um mögliche Missspezifikationen besser aufdecken zu können, wurde die Testung der Messinvarianz

zunächst für die einzelnen Konstrukte separat durchgeführt. Diese konstruktweise Untersuchung wird im Folgenden nur sehr knapp, mit Fokus auf besondere Auffälligkeiten, dargestellt. Ausführlich sind die Ergebnisse dieser Analysen in Anhang G dokumentiert. Im nächsten Schritt wurde ein gemeinsames Messmodell spezifiziert und auf Messinvarianz geprüft.

Das Vorgehen bei der Prüfung auf Messinvarianz erfolgt wie in Kapitel 4.3.4 beschrieben schrittweise durch die Testung geschachtelter Modelle mit zunehmenden Parameterrestriktionen. Dabei wurde jeweils, wie bei allen anderen Faktoren- und Strukturanalysen in dieser Arbeit, zur Skalierung der latenten Variablen die Faktorladung der je ersten Indikatorvariable auf 1 fixiert. Die Bewertung der hierarchisch geschachtelten Invarianzmodelle erfolgte anhand der beschriebenen Veränderungen des CFI_r, $RMSEA_r$ und des SRMR sowie der Gesamtbeurteilung der Passung des jeweiligen Messmodells.

Konstruktweise Invarianzprüfung – Nutzungsintention

Analog zum Vorgehen der in Kapitel 5.1.2 beschriebenen Skalenanalysen wurde zunächst das Messmodell des Konstrukts Nutzungsintention auf Messinvarianz getestet. Die Fitindikatoren dieser Messinvarianzprüfung sind in Tabelle 75 (siehe Anhang G) dargestellt. Im ersten Schritt wurde für beide Gruppen ein separates Messmodell (Modell $NI_{(V3\text{-}18,\ Teilsample)}$ und Modell $NI_{(V8\text{-}18)}$) geschätzt, das jeweils einen guten Fit aufweist. Im nächsten Schritt wurde zur Prüfung der konfiguralen Invarianz ein gemeinsames Messmodell geschätzt (Modell NI_{kInv}). Auch dieses Modell weist hinsichtlich aller Indikatoren einen mindestens akzeptablen Modellfit auf, weshalb von der Gültigkeit einer gemeinsamen Faktorstruktur in beiden untersuchten Gruppen ausgegangen werden kann und die Invarianztestung mit der Untersuchung metrischer Invarianz (Modell NI_{mInv}) fortgeführt werden konnte. Mit der schrittweisen Testung zunehmend restriktiver Modelle konnte, basierend auf den Veränderungen des CFI_r, $RMSEA_r$ und des SRMR, aufeinander aufbauend das Vorliegen metrischer, skalarer und schließlich strikter Invarianz bestätigt werden. In Tabelle 76 (siehe Anhang G) sind die zum Modell strikter Invarianz gehörigen Parameterschätzungen (Standardfehler, z-Werte, standardisierte Faktorladungen, und R^2) zu finden, die weitestgehend zwischen den beiden Gruppen vergleichbar sind.

Die Bestätigung skalarer und sogar strikter Invarianz ermöglicht die Analyse der latenten Mittelwertdifferenz zwischen VERA3- und VERA8-Lehrkräften. Im R-Zusatzpaket lavaan erfolgt dies durch die Fixierung des latenten Mittelwerts bzw. des Intercepts des latenten Konstrukts in einer der beiden betrachteten Gruppen auf null. Als Referenzgruppe, deren Intercept ebenso

wie bei allen anderen latenten Mittelwertvergleichen auf null festgesetzt wurde, dient die VERA3-Gruppe. Die Schätzung des Modells strikter Invarianz (Modell NI_{strInv}) ergibt eine auf dem 1 % Niveau signifikante standardisierte latente Mittelwertdifferenz von -0.24 (95 % KI: -0.31, -0.17). Die durchschnittliche Nutzungsintention fällt somit bei VERA8-Lehrkräften um 0.24 geringer aus als bei VERA3-Lehrkräften. Ausführlicher werden die latenten Mittelwerte bei der Analyse des gemeinsamen Messmodells aller Faktoren im nächsten Unterkapitel 5.4.3 behandelt.

Konstruktweise Invarianzprüfung – Einstellung

Die Prüfung der hierarchischen Messinvarianzstufen für das Konstrukt Einstellung ist ebenfalls im Anhang G (Tabelle 77) dokumentiert. Sowohl die separat getesteten Modelle (Modell $AE_{(V3\text{-}18,\ Teilsample)}$ und Modell $AE_{(V8\text{-}18)}$), als auch alle hierarchisch geschachtelten gemeinsamen Messmodelle weisen einen mindestens akzeptablen, hinsichtlich der meisten Fit-Kennwerte sogar einen guten Modellfit auf. Die Beurteilung der verschiedenen Invarianzstufen auf Basis von ΔCFI_r, $\Delta RMSEA_r$ und $\Delta SRMR$ bestätigt auch für das Konstrukt Einstellung das Vorliegen strikter Invarianz. Die entsprechenden Parameterschätzungen des Modells strikter Invarianz sind in Tabelle 78 (siehe Anhang G) aufgeführt. Auch für das Konstrukt Einstellung können auf Basis des Modells strikter Invarianz (Modell AE_{strInv}) latente Mittelwertunterschiede untersucht werden. Für die Einstellung zeigt sich jedoch kein signifikanter ($p > .100$) Unterschied zwischen VERA3- und VERA8-Lehrkräften. Mit einer Differenz von -0.06 (95 % KI: -0.13, -0.01) zeigt sich die Tendenz einer leicht positiveren Einstellung bei VERA3-Lehrkräften.

Konstruktweise Invarianzprüfung – Nützlichkeit und zeitliche Belastung

Wie bereits in Kapitel 5.1.2 erläutert, wurden die Konstrukte Nützlichkeit und zeitliche Belastung auch in der Gruppenanalyse aus Gründen der Identifizierbarkeit in einem gemeinsamen Messmodell geschätzt.[8] Da sich die Invarianztestung dieses Messmodells als aufwändiger erwies als die der anderen Konstrukte und die Analyse des Konstrukts zeitliche Belastung einige zusätzliche Überlegungen und Arbeitsschritte erforderte, wird über die getesteten Modelle im

[8] Die separate Invarianzprüfung für das Konstrukt Nützlichkeit liefert Belege für das Vorliegen strikter Invarianz. Die entsprechenden Kennwerte der Testung sind in Anhang G, Tabelle 79 dargestellt; die Parameterschätzungen in Tabelle 80.

Folgenden umfänglicher berichtet. Tabelle 44 dokumentiert die einzelnen Prüfschritte der Messinvarianztestung des gemeinsamen Messmodells der beiden Konstrukte.

Im ersten Schritt wurde für beide Gruppen ein separates Messmodell geschätzt (Modell $WN/ZB_{(V3\text{-}18,\ Teilsample)}$ und Modell $WN/ZB_{(V8\text{-}18)}$). Wie den Gütekriterien zu entnehmen ist, weist das Messmodell in beiden Gruppen einen guten, mindestens jedoch akzeptablen Fit auf. Im nächsten Schritt wurde zur Prüfung der konfiguralen Invarianz ein gemeinsames Messmodell geschätzt (Modell WN/ZB_{kInv}). Auch dieses Modell weist hinsichtlich aller Kennwerte einen guten Modellfit auf, weshalb von der Gültigkeit einer gemeinsamen Faktorstruktur in beiden untersuchten Gruppen ausgegangen werden kann und die Invarianztestung mit der Untersuchung metrischer Invarianz (Modell WN/ZB_{mInv}) fortgeführt werden konnte. Hierzu wurden die Faktorladungen der Indikatorvariablen jeweils zwischen den Gruppen gleichgesetzt. Im Vergleich zum unrestringierten Modell (Modell WN/ZB_{kInv}) ergibt sich durch die Restriktion der Faktorladungen keine deutliche Modellverschlechterung (ΔCFI_r = -.001, $\Delta RMSEA_r$ = .002, $\Delta SRMR$ = .003). Alle Veränderungen der relevanten Kennwerte liegen im annehmbaren Bereich. Die Gesamtbeurteilung des Modells, die zusätzlich immer zur Evaluation einer Invarianzstufe herangezogen werden sollte (Vandenberg & Lance, 2000), weist ebenfalls auf einen guten Modellfit hin.

Zur Testung skalarer Invarianz wurden zusätzlich zu den Faktorladungen die Intercepts zwischen den beiden Gruppen gleichgesetzt (Modell WN/ZB_{skInv}). Diese zusätzliche Modellrestriktion führt zu einer deutlichen Verschlechterung des Modellfits. Sowohl ΔCFI_r = -.036 als auch $\Delta RMSEA_r$ = .033 überschreiten deutlich die jeweiligen Grenzwerte einer zulässigen Verschlechterung von -.010 bzw. .015. Auch $\Delta SRMR$ = .014 überschreitet die bei der Bewertung skalarer Invarianz akzeptable Grenze von .010. Vor dem Hintergrund eines zusätzlichen, mit einem CFI_r = .949 und einem TLI_r = .943 generell unzureichenden Modellfits, muss die Hypothese skalarer Invarianz für dieses Messmodell zurückgewiesen werden.

Da jedoch für das Messmodell der Nützlichkeit in der Gruppenanalyse bereits skalare Invarianz nachgewiesen werden konnte (siehe Anhang G, Tabelle 79 und Tabelle 80), liegt die Vermutung nahe, dass die Ursache mangelnder skalarer Invarianz bei diesem Messmodell bei dem Konstrukt zeitliche Belastung liegt, welches sich bereits in Kapitel 5.1.2 als problematisch herausgestellt hat. Daher wurde als nächstes untersucht, ob durch das Freisetzen einzelner Parameter zumindest eine partielle skalare Invarianz erreicht werden kann. Hierfür wurden nacheinander die Intercepts der Indikatoritems des Konstrukts zeitliche Belastung (ZB1, ZB2, ZB3)

freigesetzt, während alle weiteren Restriktionen des Modells skalarer Invarianz (WN/ZB_{skInv}) beibehalten wurden (Modell $WN/ZB_{skInv\text{-}partiell\ a\text{-}c}$). Die Freisetzung der Intercepts von jeweils Item ZB1 und ZB2 (Modell $WN/ZB_{skInv\text{-}partiell\ a}$ und $WN/ZB_{skInv\text{-}partiell\ b}$) konnte nicht die gewünschte Modellverbesserung erzielen. Auch bei diesen Modellen liegen $\Delta CFI_r = -.022$ bzw. $\Delta CFI_r = -.026$ und $\Delta RMSEA_r = .023$ bzw. $\Delta RMSEA_r = .026$ im Vergleich zum Modell metrischer Invarianz (Modell WN/ZB_{mInv}) deutlich oberhalb einer akzeptablen Modellverschlechterung, auch wenn zumindest der Modellfit dieser beiden Modelle im akzeptablen Bereich liegt. Durch die Freisetzung des Intercepts von Item ZB3 hingegen (Modell $WN/ZB_{skInv\text{-}partiell\ c}$) kann die Hypothese partieller skalarer Invarianz bestätigt werden. Bei dieser Spezifizierung beträgt $\Delta CFI_r = -.008$, $\Delta RMSEA_r = .008$ und $\Delta SRMR = .003$, beide liegen somit unterhalb der zulässigen Grenzwerte. Auch insgesamt betrachtet erweist sich der Fit dieses Modells als gut bzw. mindestens akzeptabel.

Das Vorliegen partieller skalarer Invarianz ermöglicht zusätzlich die weitere Prüfung einer partiellen strikten Invarianz durch Gleichsetzen der Residualvarianzen (Modell $WN/ZB_{strInv\text{-}partiell}$). Da es sich hierbei basierend auf der vorangegangenen Analyse um eine Testung partieller Invarianz handelt, wurden analog zu Modell $WN/ZB_{skInv\text{-}partiell\ c}$ die Intercepts sowie Residualvarianzen des Items ZB3 frei geschätzt. Der insgesamt mindestens akzeptable Modellfit sowie die annehmbare Verschlechterung von CFI_r, $RMSEA_r$ und SRMRr ($\Delta CFI_r = -.006$, $\Delta RMSEA_r = .003$, $\Delta SRMR = .003$) sprechen dafür, dass bei diesem Messmodell von einer partiellen strikten Invarianz des Faktormodells ausgegangen werden kann. In Anhang G (siehe Tabelle 81 und Tabelle 82) sind die zum Modell partieller strikter Invarianz gehörigen Parameterschätzungen zu finden.

Das Vorliegen zumindest partieller skalarer bzw. strikter Invarianz erlaubt auch für dieses Messmodell das Schätzen latenter Mittelwertdifferenzen. Daher wurde bei der Spezifizierung des Modells partieller strikter Invarianz (Modell $WN/ZB_{strInv\text{-}partiell}$) auch die Differenz der latenten Mittelwerte geschätzt. Für das latente Konstrukt Nützlichkeit zeigt sich eine auf dem 1 % Niveau signifikante Differenz von -0.19 (95 % KI: -0.25, -0.13). Die Bewertung der Nützlichkeit ist bei VERA8-Lehrkräften somit um 0.19 niedriger als bei VERA3-Lehrkräften.

Tabelle 44: Fitstatistiken der konfirmatorischen Faktorenanalyse für das Messmodell der Konstrukte Nützlichkeit und zeitliche Belastung im Gruppenvergleich zwischen VERA3 und VERA8 – Prüfung auf Messinvarianz

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	ΔCFI_r	$\Delta RMSEA_r$	ΔSRMR
$WN/ZB_{(V3\text{-}18,\ Teilsample)}$	55.26 (.000)	19	*2.91*	**.992**	**.989**	**.034**	[.024, .045]	**.027**			
$WN/ZB_{(V8\text{-}18)}$	84.25 (.000)	19	4.43	**.972**	***.959***	*.067*	[.052, .081]	*.056*			
WN/ZB_{kInv}	137.46 (.000)	38	3.62	**.986**	**.980**	**.046**	[.038, .054]	**.035**			
WN/ZB_{mInv}	149.29 (.000)	44	3.39	**.985**	**.981**	**.044**	[.037, .052]	**.038**	*-.001*	*.002*	*.003*
WN/ZB_{skInv}	463.83 (.000)	50	9.28	.949	.943	*.077*	[.071, .084]	*.052*	-.036	.033	.014
$WN/ZB_{skInv\text{-}partiell\ a}$	321.37 (.000)	49	6.56	*.963*	*.958*	*.067*	[.060, .074]	**.042**	-.022	.023	*.004*
$WN/ZB_{skInv\text{-}partiell\ b}$	343.73 (.000)	49	7.02	*.959*	*.954*	*.070*	[.063, .077]	**.047**	-.026	.026	*.009*
$WN/ZB_{skInv\text{-}partiell\ c}$	212.84 (.000)	49	4.34	**.977**	**.974**	*.052*	[.045, .059]	**.041**	*-.008*	*.008*	*.003*
$WN/ZB_{strInv\text{-}partiell}$	264.23 (.000)	56	4.72	**.971**	**.971**	*.055*	[.049, .062]	**.047**	*-.006*	*.003*	*.006*

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben;
Modell $WN/ZB_{(V3\text{-}18,\ Teilsample)}$: Messmodell Nützlichkeit und zeitliche Belastung VERA3;
Modell $WN/ZB_{(V8\text{-}18)}$: Messmodell Nützlichkeit und zeitliche Belastung VERA8;
Modell WN/ZB_{kInv}: Konfigurale Invarianz;
Modell WN/ZB_{mInv}: Metrische Invarianz;
Modell WN/ZB_{skInv}: Skalare Invarianz;
Modell $WN/ZB_{skInv\text{-}partiell\ a\text{-}c}$: Partielle skalare Invarianz (a: Intercepts ZB1 freigeschätzt, b: Intercepts ZB2 freigeschätzt, c: Intercepts ZB3 freigeschätzt);
Modell $WN/ZB_{strInv\text{-}partiell}$: Partielle strikte Invarianz (Intercepts und Residualvarianz von Item ZB3 frei geschätzt).

Die latente Mittelwertdifferenz des Konstrukts zeitliche Belastung fällt unter der Bedingung partieller strikter Invarianz auf dem 5 % Niveau nicht signifikant aus ($p > .05$). Die Tendenz der nicht signifikanten Mittelwertdifferenz fällt mit -0.11 (95 % KI: -0.17, -0.05) negativ aus. Diese negative Tendenz widerspricht der in Kapitel 3 formulierten Erwartung, wonach VERA8-Lehrkräfte eine eher größere zeitliche Belastung empfinden. Bei der Interpretation latenter Mittelwertdifferenzen des Konstrukts zeitliche Belastung sollte jedoch berücksichtigt werden, dass ggf. die Freisetzung der Intercepts und Residuen des Items ZB3, um eine partielle skalare bzw. strikte Invarianz zu erzielen, zu einer verzerrten Schätzung von Parametern wie bspw. der latenten Intercepts geführt hat.

Zwar ist die durch Freisetzung dieses Items erzielte partielle skalare bzw. strikte Invarianz formal eine hinreichende Voraussetzung zur Analyse latenter Mittelwerte, jedoch sollten bei einer nur partiellen Invarianz immer auch inhaltliche Überlegungen berücksichtigt werden und Ungereimtheiten im Zweifelsfall kritisch hinterfragt werden. Vor dem Hintergrund, dass das Messmodell des Konstrukts zeitliche Belastung bereits in der ursprünglichen Modelltestung (siehe Kapitel 5.1.2) nicht unproblematisch war und generell die reflektive Modellierung des Konstrukts zumindest zu hinterfragen ist, liegt die Überlegung nahe, ob es unter diesen Voraussetzungen überhaupt sinnvoll erscheint, latente Mittelwertdifferenzen zu analysieren, oder ob die Mittelwerte dieses Konstrukts besser nur auf manifester Ebene betrachtet werden sollten.

Auch ein Blick auf die Mittelwerte der Indikatorvariablen (siehe Tabelle 43), welche den unstandardisierten, nicht restringierten Intercepts im Modell konfiguraler und metrischer Invarianz (Modell WN/ZB_{kInv} und WN/ZB_{mInv}) entsprechen, offenbart eine nicht unwesentliche Ambivalenz in den Einschätzungen der Lehrkräfte beider Gruppen hinsichtlich der verschiedenen Aspekte zeitlicher Belastung. Wie bereits in Kapitel 5.4.1 beschrieben, schätzen VERA3-Lehrkräfte die Belastung durch die Vorbereitung der Vergleichsarbeiten (Item ZB1) höher ein ($M = 2.71$, $SD = 0.90$) als die VERA8-Lehrkräfte ($M = 2.44$, $SD = 1.04$). Item ZB3 hingegen wird von den VERA8-Lehrkräften negativer bewertet ($M = 3.88$, $SD = 0.88$) als durch die Lehrkräfte der VERA3-Gruppe ($M = 3.40$, $SD = 0.78$) und zeigt einen höheren empfundenen Arbeitsaufwand der Auswertung bei Lehrkräften in der 8. Klasse an. Die Unterschiede bei Item ZB2 und somit der Bewertung des Durchführungsaufwandes hingegen sind nur sehr gering (VERA3: $M = 3.04$, $SD = 0.68$; VERA8: $M = 3.02$, $SD = 0.85$).

Dies alles spricht dafür, dass eine Gleichsetzung der Intercepts der Indikatorvariablen dieses Konstrukts die Daten nicht adäquat abbildet. Durch Freisetzung von Item ZB3 kann zwar im Hinblick auf den Modellfit von einer partiellen skalaren bzw. strikten Invarianz ausgegangen werden, jedoch bleibt die Problematik erzwungener Gleichsetzung bei Item ZB1 bestehen. Das Freisetzen beider Items (ZB1 und ZB3) könnte auf Basis von Modell $WN/ZB_{skInv\text{-}partiell\ c}$ den Modellfit eines Modells partieller skalarer Invarianz noch weiter verbessern (χ_r^2 (48) = 179.63, $p < .001$, $CFI_r = .982$, $TLI_r = .979$, $RMSEA_r = .045$, SRMR = .039). Eine Freisetzung der Intercepts beider betroffener Items (ZB1 und ZB3) würde jedoch den Richtlinien der Spezifizierung eines zulässigen partiell invarianten Messmodells, bei dem die Mehrheit der Indikatoritems eines Konstrukts Invarianz aufweisen sollten, widersprechen und wiederum auch nicht zu sinnvoll interpretierbaren Ergebnissen führen.

Auf Grundlage dieser Ausführungen erschien es nicht sinnvoll, für das Konstrukt zeitliche Belastung latente Mittelwertdifferenzen zu untersuchen, weshalb in den folgenden Abschnitten bei diesem Konstrukt darauf verzichtet wurde und für eine inhaltliche Interpretation ergänzend nur manifeste Mittelwertunterschiede herangezogen wurden.

CFA mit allen Konstrukten

Unter Berücksichtigung der Erkenntnisse der konstruktweisen Testung der Messinvarianz wurde ein gemeinsames Messmodell für alle untersuchten Konstrukte aufgestellt und auf Messinvarianz geprüft. In Tabelle 45 sind die Fitstatistiken der im Folgenden beschriebenen Modelle wiedergegeben. Wie gehabt wurde im ersten Analyseschritt für beide Gruppen ein separates Messmodell mit identischen Spezifizierungen geschätzt (Modell $GroupCFA_{(V3\text{-}18,\ Teilsample)}$ und Modell $GroupCFA_{(V8\text{-}18)}$). Wie den Gütekriterien zu entnehmen ist, beschreibt das Modell die Faktorstruktur der zugrundeliegenden Daten in beiden Gruppen vergleichbar gut.

Im zweiten Schritt wurde zur Testung der einheitlichen Faktorstruktur ein gemeinsames Messmodell beider Gruppen geschätzt (Modell $GroupCFA_{kInv}$). Dieses Modell konfiguraler Invarianz weist über alle Fitstatistiken hinweg einen guten Modellfit auf ($\chi_r^2(258) = 742.54$, $p < .001$, $\chi^2/df = 2.88$, $CFI_r = .983$, $TLI_r = .980$, $RMSEA_r = .040$, SRMR = .034), sodass die Prüfung weiterer Invarianzstufen fortgesetzt werden konnte. Wird die Restriktion gleicher Faktorladungen zwischen den Gruppen hinzugefügt (Modell $GroupCFA_{mInv}$, metrische

Invarianz), führt dies zu einer nur unwesentlichen Verschlechterung des Modellfits gegenüber dem weniger restriktiven Modell $\text{GroupCFA}_{\text{kInv}}$ (ΔCFI_r = -.001, ΔRMSEA_r = .000, ΔSRMR = .003). Die Verschlechterung des CFI_r und des SRMR liegen im annehmbaren Bereich, der RMSEA_r bleibt unverändert. Auch das Modell im Gesamten beschreibt die Daten beider Gruppen entsprechend gut ($\chi_r^2(272) = 770.26$, $p < .001$, $\chi^2/\text{df} = 2.83$, CFI_r = .982, TLI_r = .980, RMSEA_r = .040, SRMR = .037). Die Minimalvoraussetzung zur gruppenübergreifenden Analyse von Strukturbeziehungen ist somit gegeben.

Der nächste Prüfschritt bezieht sich auf die Testung skalarer Invarianz durch die Annahme gleicher Intercepts in beiden Gruppen. Entsprechend den Erkenntnissen aus den Vorarbeiten im vorangegangenen Abschnitt zu den Problemen des Konstrukts zeitliche Belastung wurde hierfür direkt ein Modell partieller skalarer Invarianz spezifiziert, in dem die Intercepts aller Indikatorvariablen außer die des Items ZB3 des Konstrukts zeitliche Belastung gleichgesetzt wurden (Modell $\text{GroupCFA}_{\text{skInv-partiell}}$). Durch die Annahme (partieller) Äquivalenz der Indikatorkonstanten verschlechtert sich der Modellfit im Vergleich zum Modell metrischer Invarianz (Modell $\text{GroupCFA}_{\text{mInv}}$) erneut nur geringfügig (ΔCFI_r = -.002, ΔRMSEA_r = .002, ΔSRMR = .001). Auch auf Basis der Gesamtbeurteilung des Modellfits ($\chi_r^2(285) = 867.19$, $p < .001$, $\chi^2/df = 3.04$, CFI_r = .980, TLI_r = .978, RMSEA_r = .042, SRMR = .038) kann vom Vorliegen partieller skalarer Invarianz ausgegangen werden, wodurch die formale Voraussetzung zur Analyse latenter Mittelwertdifferenzen erfüllt ist.

Der Vollständigkeit halber sei hier noch das Ergebnis der Prüfung vollständiger skalarer Invarianz berichtet: Das entsprechende Modell weist zwar einen guten Fit auf ($\chi_r^2(286) = 1\,052.78$, $p < .001$, $\chi^2/df = 3.68$, CFI_r = .973, TLI_r = .971, RMSEA_r = .048, SRMR = .035) und auch die Bedingung $\Delta\text{CFI}_r \leq$ -.010 ist mit ΔCFI_r = -.009 noch erfüllt, jedoch verliert die Varianz des latenten Faktors zeitliche Belastung an Signifikanz (Var(η) = 0.080, p = .04), was auf eine Fehlspezifikation des Modells hinweist, die bereits im vorangegangenen Abschnitt ausführlich erläutert wurde. Aufgrund der Mängel des Modells skalarer Invarianz wurde die Invarianzprüfung auf Basis des Modells partieller skalarer Invarianz (Modell $\text{GroupCFA}_{\text{skInv-partiell}}$) fortgesetzt.

Zur Prüfung (partieller) strikter Invarianz (Modell $\text{GroupCFA}_{\text{strInv-partiell a}}$) wurden, neben den bereits in Modell $\text{GroupCFA}_{\text{skInv-partiell}}$ spezifizierten Restriktionen, zusätzlich die Fehlervarianzen der Indikatorvariablen gleichgesetzt, nur für den Indikator ZB3 wurde auch dieser Wert,

ebenso wie die Intercepts, jeweils frei geschätzt. Der Blick auf die Fitstatistiken und den Umfang der Verschlechterung des Modellfits bestätigt das Vorliegen partieller strikter Invarianz: Der Gesamtmodellfit verbleibt über alle Kennwerte hinweg gut ($\chi_r^2(302) = 957.66$, $p < .001$, $\chi^2/df = 3.17$, $CFI_r = .977$, $TLI_r = .976$, $RMSEA_r = .043$, $SRMR = .042$). Die Abnahme der Modellpassung im Vergleich zu Modell $GroupCFA_{skInv\text{-}partiell}$ liegt mit $\Delta CFI_r = -.003$, $\Delta RMSEA_r = .001$, $\Delta SRMR = .004$ im akzeptablen Bereich.

Bei der Prüfung skalarer und strikter Invarianz wurden die Gruppenunterschiede der latenten Mittelwerte durch Fixierung der Intercepts der latenten Konstrukte auf null mitgeschätzt. Da jedoch, wie bereits dargelegt, der latente Mittelwertvergleich für das Konstrukt zeitliche Belastung aus inhaltlichen Gründen wenig sinnvoll erscheint, wurde ein weiteres Modell geschätzt (Modell $GroupCFA_{strInv\text{-}partiell\ b}$), in welchem die Intercepts des Konstrukts zeitliche Belastung in beiden Gruppen auf null fixiert werden, alle weiteren Parameterschätzungen erfolgen analog zu Modell $GroupCFA_{strInv\text{-}partiell\ a}$. Dieses Modell $GroupCFA_{strInv\text{-}partiell\ b}$ weist gegenüber Modell $GroupCFA_{strInv\text{-}partiell\ a}$, ohne diese zusätzliche Restriktion, nur eine minimale Verschlechterung auf und wurde daher auch aufgrund inhaltlicher Plausibilitätsüberlegungen beibehalten und zur Analyse latenter Mittelwertdifferenzen und zur auf der Messinvarianzprüfung aufbauenden Analyse von Strukturbeziehungen genutzt.

Tabelle 45: Fitstatistiken der konfirmatorischen Faktorenanalyse für das Messmodell mit allen Konstrukten im Gruppenvergleich zwischen VERA3 und VERA8 – Prüfung auf Messinvarianz

Modell	χ_r^2 (*p*-Wert)	df	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	ΔCFI_r	$\Delta RMSEA_r$	ΔSRMR
$GroupCFA_{(V3\text{-}18,\ Teilsample)}$	463.84 (.000)	129	3.60	**.983**	**.980**	**.040**	[.036, .044]	**.033**			
$GroupCFA_{(V8\text{-}18)}$	277.54 (.000)	129	*2.15*	**.982**	**.979**	**.041**	[.035, .048]	**.037**			
$GroupCFA_{kInv}$	742.54 (.000)	258	*2.88*	**.983**	**.980**	**.040**	[.037, .044]	**.034**			
$GroupCFA_{mInv}$	770.26 (.000)	272	*2.83*	**.982**	**.980**	**.040**	[.036, .043]	**.037**	-.001	.000	.003
$GroupCFA_{skInv\text{-}partiell}$	867.19 (.000)	285	3.04	**.980**	**.978**	**.042**	[.039, .045]	**.038**	-.002	.002	.002
$GroupCFA_{strInv\text{-}partiell\ a}$	957.66 (.000)	302	3.17	**.977**	**.976**	**.043**	[.040, .046]	**.042**	-.003	.001	.004
$GroupCFA_{strInv\text{-}partiell\ b}$	962.63 (.000)	303	3.18	**.976**	**.976**	**.043**	[.040, .047]	**.043**			

Anmerkungen. VERA8 2018: *N* = 782; VERA3 2018-Teilstichprobe: *N* = 1 918; gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben;
Modell $GroupCFA_{(V3\text{-}18,\ Teilsample)}$: Messmodell Nützlichkeit und zeitliche Belastung VERA3;
Modell $GroupCFA_{(V8\text{-}18)}$: Messmodell Nützlichkeit und zeitliche Belastung VERA8;
Modell $GroupCFA_{kInv}$: Konfigurale Invarianz;
Modell $GroupCFA_{mInv}$: Metrische Invarianz;
Modell $GroupCFA_{skInv\text{-}partiell}$: Partielle skalare Invarianz (3c: Intercepts ZB3 freigeschätzt);
Modell $GroupCFA_{strInv\text{-}partiell\ a}$: Partielle strikte Invarianz (Intercepts und Residualvarianz von Item ZB3 frei geschätzt);
Modell $GroupCFA_{strInv\text{-}partiell\ b}$: Partielle strikte Invarianz (Intercepts und Residualvarianz von Item ZB3 frei geschätzt), Intercepts des Konstrukts zeitliche Belastung in beiden Gruppen auf 0 fixiert.

In Tabelle 46 sind Standardfehler, z-Werte, standardisierte Faktorladungen und R^2 der Indikatoritems des finalen Modells $GroupCFA_{strInv\text{-}partiell\ b}$ zusammengefasst. Die Schätzungen der Intercepts und Messfehlervarianzen für Modell $GroupCFA_{strInv\text{-}partiell\ b}$ sind in Anhang H (siehe Tabelle 83 und Tabelle 84) zu finden.

Tabelle 46: Standardisierte Faktorladungen, aufgeklärte Varianz der Indikatorvariablen, Standardfehler und z-Werte in Modell $GroupCFA_{strInv\text{-}partiell\ b}$ (partielle strikte Invarianz) im Gruppenvergleich zwischen VERA3 und VERA8

		VERA3		VERA8			
Konstrukt	Item	λ^s_{ij}	R^2	λ^s_{ij}	R^2	*S. E.* [c]	*z*-Wert [c]
Nutzungs-intention [a]	NI1	.888	.789	.893	.797	.000	
	NI2	.846	.716	.852	.726	.013	71.547
	NI3	.837	.701	.843	.711	.014	64.089
	NI4	.703	.494	.711	.506	.018	44.990
	NI5	.644	.414	.653	.426	.022	34.413
Einstellung [a]	AE1	.929	.864	.936	.876	.000	
	AE2	.829	.687	.842	.710	.013	63.714
	AE3	.907	.823	.915	.838	.010	97.501
	AE4	.887	.786	.896	.803	.011	88.192
	AE5	.815	.664	.829	.687	.014	65.008
Nützlichkeit [a]	WN1	.824	.679	.836	.699	.000	
	WN2	.767	.589	.781	.610	.021	46.163
	WN3	.826	.683	.838	.702	.020	54.245
	WN4	.724	.524	.739	.546	.022	41.558
	WN5	.777	.604	.791	.626	.022	46.335
Zeitliche Belastung [b]	ZB1	.393	.155	.489	.239	.000	
	ZB2	.685	.469	.776	.602	.091	14.381
	ZB3	.535	.286	.590	.349	.125	8.920

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; [c] Durch die gesetzten Restriktionen sind *S. E.* und *z*-Werte in beiden Gruppen identisch; Alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Die Faktorladungen der Items der Konstrukte Nutzungsintention, Einstellung und Nützlichkeit fallen in beiden Gruppen hoch aus und liegen zwischen .64 und .94. Insgesamt sind die Faktorladungen für die VERA8-Gruppe etwas höher, der Unterschied beträgt für die betrachteten

drei Konstrukte zwischen .005 und .015. Etwas stärker fällt die Differenz in der Höhe der Faktorladungen dagegen für die Indikatorvariablen des Konstrukts zeitliche Belastung aus. Hier fallen die Faktorladungen ebenfalls in der VERA8-Gruppe größer aus, der Unterschied liegt hier zwischen .055 und .096. Ein ähnliches Bild zeigt sich bei der Varianzaufklärung der Indikatorvariablen. Sie ist ebenfalls in der VERA8-Gruppe etwas höher, wobei die Differenz bei den Items des Konstrukts zeitliche Belastung zwischen .063 und .133 liegt und bei den übrigen drei Konstrukten nur zwischen .008 und .023.

5.4.3. (Latente) Konstruktmittelwerte und Faktorkorrelationen

Tabelle 47 enthält die auf Basis von Modell $\text{GroupCFA}_{\text{strInv-partiell b}}$ geschätzten Faktorkorrelationen und latenten Konstruktmittelwerte sowie die manifesten Konstruktmittelwerte. Die latenten Faktormittelwerte der Konstrukte Nutzungsintention, Einstellung und Nützlichkeit offenbaren für dieses Modell $\text{GroupCFA}_{\text{strInv-partiell b}}$ ein erwartungskonformes Bild. Die latenten Mittelwerte sind in Tabelle 47 einschließlich ihrer 95 %-Konfidenzintervalle wiedergegeben. Wie an dem negativen Vorzeichen zu erkennen ist, fallen die latenten Mittelwerte in der Gruppe der VERA8-Lehrkräfte über alle drei Konstrukte hinweg signifikant niedriger aus als unter den VERA3-Lehrkräften. Die größte Mittelwertdifferenz liegt dabei mit -.26 bei dem Konstrukt Nutzungsintention, gefolgt vom Gruppenunterschied der wahrgenommenen Nützlichkeit von -.22. Beide Mittelwertunterschiede erweisen sich auf dem 1 % Niveau als signifikant. Die Differenz des Einstellungskonstrukts weist die gleiche Tendenz auf, fällt jedoch mit -.14 etwas geringer aus (p = .001). Diese Ergebnisse spiegeln somit auch im latenten Gruppenmodell wider, was die manifesten Skalenmittelwerte andeuten. Auch hier zeigt sich bei den VERA8-Lehrkräften eine negativere Wahrnehmung hinsichtlich der einzelnen Konstrukte.

Dies ist auch beim Konstrukt zeitliche Belastung der Fall, dessen Mittelwertunterschiede aus den genannten messtheoretischen und inhaltlichen Gründen nicht auf latenter Ebene analysiert wurden. Dennoch lohnt sich auch hier ein Blick auf die manifesten Mittelwerte. Auf Konstruktebene unterscheiden sich diese nur geringfügig. Der Mittelwert des Konstrukts liegt in der VERA3-Gruppe bei 3.05, in der VERA8-Gruppe bei 3.11. In der Durchführung eines Welch-t-Tests erweist sich dieser Unterschied auf dem 5 % Niveau als signifikant (t(1 216.70) = -2.21, p = .027) (siehe Tabelle 48). In Anbetracht der negativen Polung der

Skala bedeutet dies auch für dieses Konstrukt eine insgesamt leicht schlechtere Bewertung in der VERA8-Gruppe. Auffällig ist bei diesem Konstrukt zudem die, im Vergleich zu den anderen Konstrukten, geringere Streuung von 0.58 bzw. 0.68.

Tabelle 47: Korrelationen der latenten Faktoren (95 % KI), Mittelwerte der latenten Faktoren (95 % KI) auf Basis von Modell GroupCFA$_{strInv\text{-}partiell\ b}$ (partiell strikte Invarianz) (vollständig standardisierte Lösung) sowie manifeste Konstruktmittelwerte (Standardabweichungen)

		Faktorkorrelationen			Faktor-mittelwerte (η_i)	Manifeste Mittelwerte (M)
		NI	AE	ZB		
VERA3	NI [a]				0.0 (–)	2.36 (0.69)
	AE [a]	.74 (.71, .77)			0.0 (–)	2.42 (0.80)
	ZB [b]	-.23 (-.25, -.21)	-.33 (-.35, -.30)		0.0 (–)	3.05 (0.58)
	WN [a]	.81 (.77, .84)	.81 (.78, .84)	-.30 (-.32, -.27)	0.0 (–)	2.45 (0.74)
VERA8	NI [a]				-0.26 (-0.32, -0.19)	2.20 (0.71)
	AE [a]	.79 (.74, .84)			-0.14* (-0.21, -0.07)	2.34 (0.84)
	ZB [b]	-.37 (-.41, -.33)	-.47 (-.51, -.42)		n.b.	3.11 (0.68)
	WN [a]	.87 (.82, .92)	.86 (.81, .91)	-.43 (-.47, -.39)	-0.22 (-0.28, -0.16)	2.31 (0.78)

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; NI: Nutzungsintention, AE: Einstellung, ZB: Zeitliche Belastung, WN: Nützlichkeit; V3: VERA3, V8: VERA8; n.b.: nicht berechnet; [a] 4-stufige positiv gepolte Antwortskala; [b] 5-stufige negativ gepolte Antwortskala; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$), außer *$p = .001$.

Bei der Interpretation von Mittelwertunterschieden sollte jedoch bei diesem Konstrukt differenzierter vorgegangen werden, da ein Blick auf die Mittelwerte der Indikatorvariablen (siehe Tabelle 48), wie auch bereits in Kapitel 5.4.1 angeschnitten, offenlegt, dass diese zwischen den Gruppen sehr unterschiedliche Tendenzen aufweisen. Wie ebenfalls Tabelle 48 zu entnehmen ist, erweisen sich die Gruppenunterschiede bei den Items ZB1 und ZB3 auf dem 1 %

Niveau als signifikant, die deutlich kleinere Mittelwertdifferenz zwischen den Gruppen bei Item ZB2 hingegen nicht. Entsprechend zeigt die Betrachtung der Effektstärken für Item ZB2 mit $d = 0.03$ keinen Effekt, während bei ZB1 zumindest ein kleiner ($d = 0.27$) und bei ZB3 mit $d = -0.56$ ein mittlerer Gruppeneffekt erkennbar ist. Der Vorbereitungsaufwand der Vergleichsarbeiten erweist sich somit in der Wahrnehmung der Grundschullehrkräfte als größer als für Sekundarschullehrkräfte. Den Aufwand der Auswertung dagegen empfinden Lehrkräfte bei der VERA8-Testung als höher als bei VERA3. Den Durchführungsaufwand bewerten beide Gruppen im Mittel als relativ neutral bzw. angemessen.

Tabelle 48: Signifikanztests zu Gruppenunterschieden zwischen VERA3 und VERA8

Konstrukt/ Item	Gruppe	Gruppenstatistiken			Levene-Test		t-Test auf Mittelwertgleichheit [a]			Effekt-stärke
		n	*M*	*SD*	*F*	*p*	*t*	*df*	*p*	*d*
ZB	VERA3	1 885	3.05	0.58	37.69	< .001	-2.21	1 216.7	.027	0.10
	VERA8	755	3.11	0.68						
ZB1	VERA3	1 900	2.71	0.90	53.26	< .001	6.37	1 238.7	< .001	0.27
	VERA8	759	2.44	1.04						
ZB2	VERA3	1 906	3.04	0.68	34.28	< .001	0.63	1 191.6	.528	0.03
	VERA8	770	3.02	0.85						
ZB3	VERA3	1 904	3.40	0.78	30.35	< .001	-13.06	1 307.4	< .001	-0.56
	VERA8	778	3.88	0.88						

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; ZB: zeitliche Belastung;
[a] Welch-t-Test für ungleiche Varianzen, da basierend auf dem Levene-Test die Hypothese der Varianzhomogenität verworfen wird.

Mit Blick auf die Faktorkorrelationen (siehe Tabelle 47) stellen sich zwischen den Gruppen vergleichbare Tendenzen heraus: Zunächst besitzen die einzelnen Korrelationen jeweils in beiden Gruppen dasselbe Vorzeichen. Betrachtet man des Weiteren, wie sich die Stärke der einzelnen Zusammenhänge innerhalb der Gruppen zueinander verhält, ergeben sich ebenfalls jeweils ähnliche Muster. Die Stärke der einzelnen Korrelationen unterscheidet sich jedoch zwischen den Gruppen. Alle Zusammenhänge fallen in der Gruppe der VERA8-Lehrkräfte

stärker aus. In der Gruppe der VERA3-Lehrkräfte liegt der Betrag der Korrelationskoeffizenten zwischen .23 und .81, und in der VERA8-Gruppe zwischen .37 und .87. Besonders groß ist der Unterschied bei Korrelationen mit dem Konstrukt zeitliche Belastung. Hier unterscheidet sich die Höhe der Korrelation bis zu .14., die Interkorrelation zwischen zeitlicher Belastung und Einstellung liegt bspw. in der VERA3-Gruppe bei r = -.33 und in der VERA8-Gruppe bei r = -.47. Basierend auf den fehlenden Überschneidungen der 95 %-Konfidenzintervalle dieser Korrelationen (ZB ↔ NI, ZB ↔ AE, ZB ↔ WN), kann auf einen signifikanten Unterschied dieser Faktorkorrelationen zwischen den Gruppen geschlossen werden.

Insgesamt liefert die Analyse der korrelativen Zusammenhänge erste Hinweise darauf, dass neben dem Messmodell auch die Struktur der Faktorbeziehungen zwischen den betrachteten Gruppen vergleichbar ist, aber möglicherweise, wie vermutet, in der VERA8-Gruppe ein stärkerer Einfluss der zeitlichen Belastung besteht. Dies wird im nächsten Abschnitt bei der Analyse der Strukturbeziehungen noch weiter untersucht werden.

5.4.4. Analyse des Strukturmodells

Auf Grundlage der nachgewiesenen partiellen strikten Invarianz des Messmodells (Modell $\text{GroupCFA}_{\text{strInv-partiell b}}$) konnten im nächsten Schritt mögliche Unterschiede in den Strukturbeziehungen des angepassten Kausalmodells untersucht werden (siehe Kapitel 5.2, Modell 2). Wie bei den vorangegangenen Testungen der Kausalbeziehungen in den Kapiteln 5.1.3, 5.2 und 5.3, erfolgte die Schätzung des Strukturmodells in zwei Schritten: Zunächst wurde kein direkter Pfad zwischen dem Konstrukt Nützlichkeit und Nutzungsintention spezifiziert, im nächsten Analyseschritt wurde dieser zugelassen. Gemäß diesem Vorgehen wurden basierend auf Messmodell $\text{GroupCFA}_{\text{strInv-partiell b}}$ die Kausalbeziehungen zwischen den Konstrukten spezifiziert und die entsprechenden Modelle geschätzt (Modell 2 $\text{Group}_{(\text{PKf})}$ und Modell 2_{dP} $\text{Group}_{(\text{PKf})}$). Zusätzlich wurden beide Modelle mit zwischen den Gruppen gleichgesetzten Pfadkoeffizienten geschätzt (Modell 2 $\text{Group}_{(\text{PKg})}$ und Modell 2_{dP} $\text{Group}_{(\text{PKg})}$).

Tabelle 49 enthält die Fitstatistiken dieser Modellschätzungen. Wie bei der ursprünglichen Testung des angepassten Kausalmodells (siehe Kapitel 5.2 und 5.3), weist auch das Gruppenmodell (Modell 2 $\text{Group}_{(\text{PKf})}$) zunächst einen akzeptablen Modellfit auf, der sich zudem durch die Spezifikation des zusätzlichen Pfades (Modell 2_{dP} $\text{Group}_{(\text{PKf})}$) nochmals deutlich

verbessert. Durch die Gleichsetzung der Pfadkoeffizienten (Modell 2 $Group_{(PKg)}$ und 2_{dP} $Group_{(PKg)}$) zeigt sich mit $\Delta CFI_r = .000$ jeweils keine substanzielle Verschlechterung der Modellpassung, weshalb die Modelle 2 $Group_{(PKg)}$ und 2_{dP} $Group_{(PKg)}$ als restriktivere Alternativen akzeptiert und auf deren Basis die Parameterschätzungen analysiert werden. Insgesamt weisen beide Modelle mit dem zusätzlichen direkten Pfad zwischen Nützlichkeit und Nutzungsintention (Modell 2_{dP} $Group_{(PKf)}$ und 2_{dP} $Group_{(PKg)}$) einen mit Messmodell 4b vergleichbaren Modellfit auf. Der gleichbleibend gute Modellfit bei gleichgesetzten Pfadkoeffizienten spricht dafür, dass sich die Struktur der Pfadkoeffizienten zwischen den Gruppen nicht bedeutsam unterscheidet und das postulierte Kausalmodell für beide Gruppen Gültigkeit hat. Dies bestätigt sich auch bei der Betrachtung der Pfadkoeffizienten.

Tabelle 49: Fitstatistiken der Strukturgleichungsmodelle des Gruppenvergleichs auf Basis von Messmodell $GroupCFA_{strInv\text{-}partiell\ b}$ der Invarianzanalyse und Strukturmodell 2 und 2_{dP}

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	AIC
2 $Group_{(PKf)}$	1 335.46 (.000)	307	4.35	*.963*	*.963*	*.054*	[.051, .057]	*.058*	88 756
2_{dP} $Group_{(PKf)}$	965.43 (.000)	305	3.17	**.976**	**.976**	**.043**	[.040, .046]	**.043**	88 325
2 $Group_{(PKg)}$	1 339.77 (.000)	311	4.31	*.963*	*.964*	*.054*	[.051, .057]	*.059*	88 752
2_{dP} $Group_{(PKg)}$	970.70 (.000)	310	3.13	**.976**	**.977**	**.043**	[.040, .046]	**.044**	88 321

Anmerkungen. VERA8 2018: *N* = 782; VERA3 2018-Teilstichprobe: *N* = 1 918; gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben; Alle Strukturmodelle basierend auf Messmodell $GroupCFA_{strInv\text{-}partiell\ b}$: Partielle strikte Invarianz (Intercepts und Residualvarianz von Item ZB3 frei geschätzt), Intercepts des Konstrukts zeitliche Belastung in beiden Gruppen auf 0 fixiert;
Modell 2 $Group_{(PKf)}$: Strukturmodell 2 (kein direkter Pfad von Nützlichkeit und Nutzungsintention), Pfadkoeffizienten frei geschätzt;
Modell 2_{dP} $Group_{(PKf)}$: Strukturmodell 2_{dP} (direkter Pfad von Nützlichkeit und Nutzungsintention), Pfadkoeffizienten frei geschätzt;
Modell 2 $Group_{(PKg)}$: Strukturmodell 2 (kein direkter Pfad von Nützlichkeit und Nutzungsintention), Pfadkoeffizienten gleichgesetzt;
Modell 2_{dP} $Group_{(PKg)}$: Strukturmodell 2_{dP} (direkter Pfad von Nützlichkeit und Nutzungsintention), Pfadkoeffizienten gleichgesetzt.

Abbildung 21 visualisiert zunächst das Strukturmodell 2 $Group_{(PKg)}$ ohne direkten Pfad zwischen Nützlichkeit und Nutzungsintention. Dargestellt sind neben den Fitstatistiken die standardisierten Faktorladungen und die Varianzaufklärung je Konstrukt (R^2) im Gruppenvergleich. Alle weiteren Parameterschätzungen zu Modell 2 $Group_{(PKg)}$ sind in Anhang I (siehe Tabelle 85, Tabelle 86 und Tabelle 87) zu finden. Der erste Wert bezieht sich jeweils auf die

Gruppe der VERA3-Lehrkräfte, der zweite Wert auf die VERA8-Gruppe. Der Blick auf die Effektschätzungen verdeutlicht, dass die Schätzung der Pfadkoeffizienten in beiden Gruppen eine weitgehend übereinstimmende Kausalstruktur widerspiegelt. Alle Pfadkoeffizienten erweisen sich dabei als signifikant ($p < .001$). Die Vorzeichen der Pfadkoeffizienten sind in den Gruppen identisch, und auch die Höhe der Effekte unterscheidet sich nur schwach und ist mit den Effekten der zuvor mit VERA3-Daten geschätzten Modelle vergleichbar. Die standardisierten Pfadkoeffizienten fallen für die VERA8-Gruppe alle etwas höher aus, jedoch beträgt die Differenz im Betrag jeweils nur .02 bzw. .03. Der größte Unterschied zeigt sich bei der Beziehung zwischen zeitlicher Belastung und wahrgenommener Nützlichkeit, hier liegt der Effekt in der VERA3-Gruppe bei (γ = -.32), in der VERA8-Gruppe bei (γ = -.40).

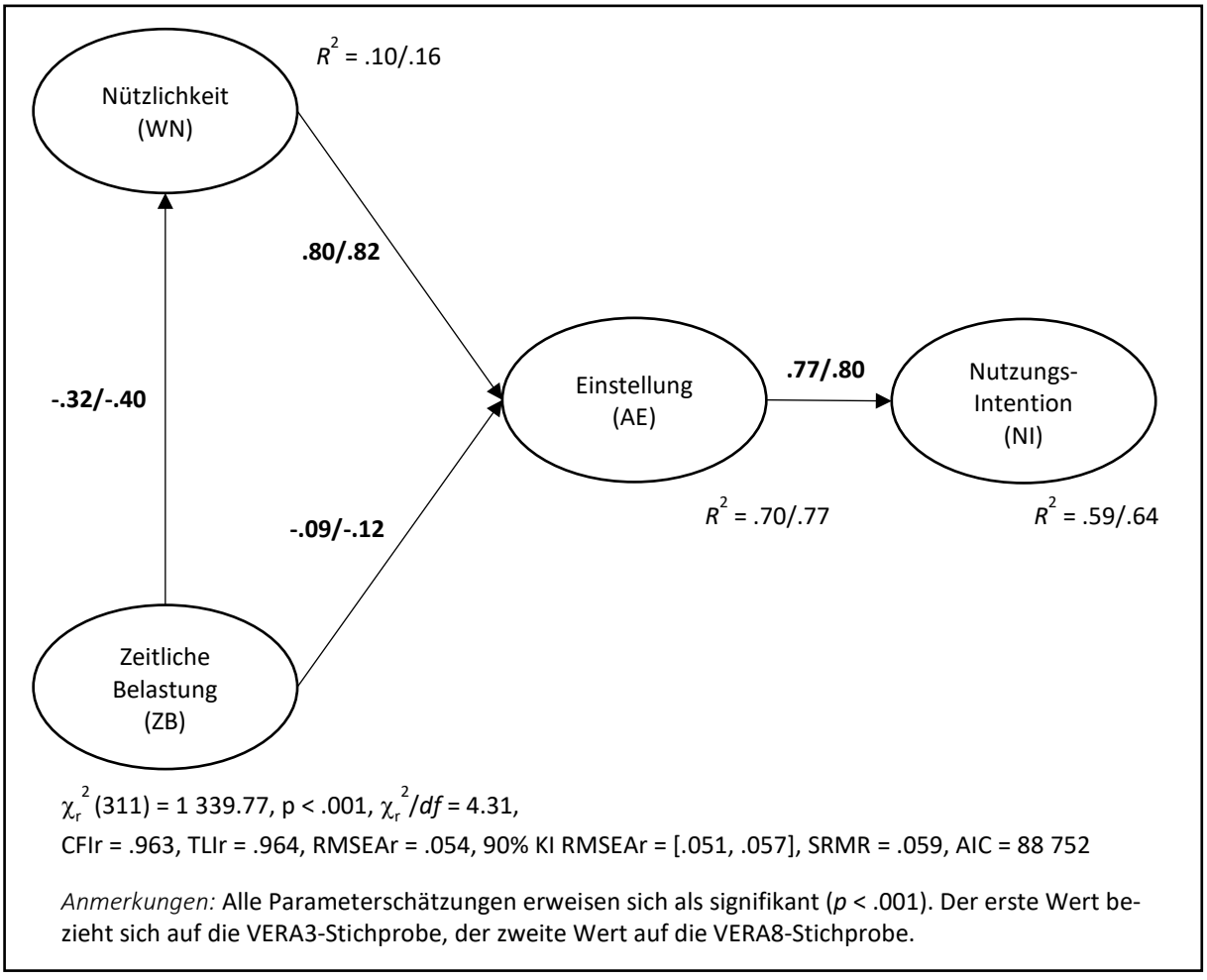


Abbildung 21: Modell 2 Group$_{(PKg)}$ – Fitstatistiken und Parameterschätzungen im Gruppenvergleich VERA3 und VERA8 (vollständig standardisierte Lösung).

Die größere Bedeutung des zeitlichen Aspekts für VERA8-Lehrkräfte spiegelt sich auch in der größeren Varianzaufklärung der wahrgenommenen Nützlichkeit wider. Erklärt die

empfundene zeitliche Belastung in der VERA3-Gruppe nur 10 % der Varianz des Konstrukts Nützlichkeit, liegt die Varianzaufklärung in der VERA8-Gruppe bei 16 %. Auch bei den anderen endogenen latenten Faktoren fällt die Varianzaufklärung in der VERA8-Gruppe höher aus als in der VERA3-Gruppe.

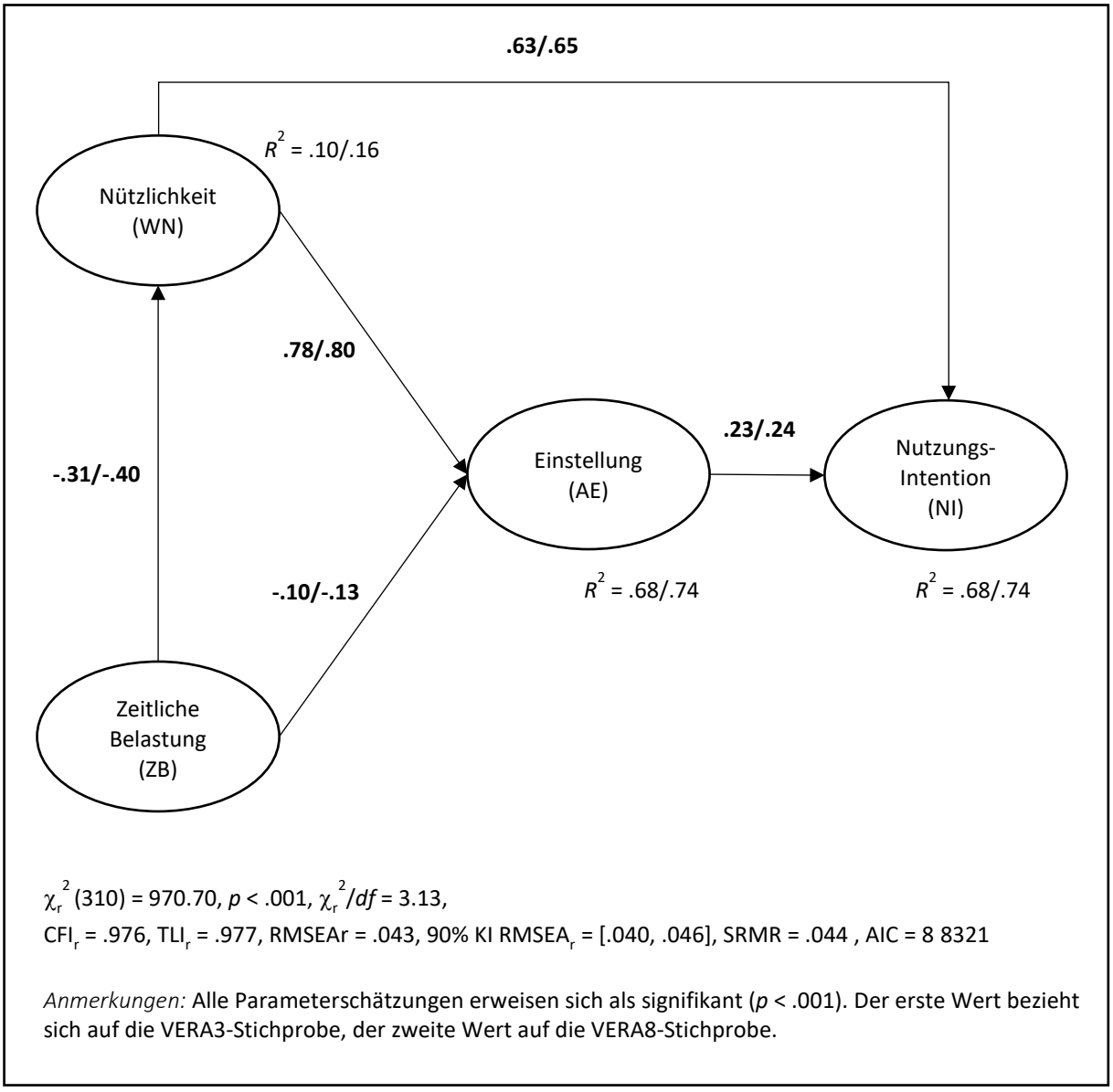


Abbildung 22: Modell 2_{dP} $Group_{(PKg)}$ – Fitstatistiken und Parameterschätzungen im Gruppenvergleich VERA3 und VERA8 (vollständig standardisierte Lösung)

Auch in Modell 2_{dP} $Group_{(PKg)}$ (siehe Abbildung 22) bleiben alle Effekte auf dem 1 % Niveau signifikant. Durch die Spezifikation des zusätzlichen direkten Pfades verändern sich hier die Effekte der exogenen Variablen zeitliche Belastung in beiden Gruppen kaum. Auch der Effekt

der wahrgenommenen Nützlichkeit auf das Konstrukt Einstellung nimmt nur geringfügig ab. Wie bereits im VERA3-Strukturmodell (siehe Modell 2_{dP}, Kapitel 5.2) verändern sich vor allem die Effekte, die auf die Nutzungsintention wirken. Der Pfadkoeffizient der Einstellung auf die Nutzungsintention sinkt dementsprechend auf β_{V3} = .23 bzw. β_{V8} = .24, der Effekt zwischen Nützlichkeit und Nutzungsintention liegt dagegen bei β_{V3} = .63 bzw. β_{V8} = .65. Insgesamt treten auch in Modell 2_{dP} $Group_{(PKg)}$ keine substanziellen Unterschiede in den Effekten zwischen den Gruppen hervor. Hinsichtlich der Varianzaufklärung zeigt sich ebenfalls ein zu Modell 2 $Group_{(PKg)}$ vergleichbares Bild einer höheren Varianzaufklärung der endogenen latenten Faktoren in der VERA8-Gruppe. Die Varianzaufklärung der Nutzungsintention nimmt im Vergleich zu Modell 2 $Group_{(PKg)}$ in beiden Gruppen zu, für das Konstrukt Einstellung nimmt sie jeweils marginal ab. Alle weiteren Parameterschätzungen zu Modell 2_{dP} $Group_{(PKg)}$ sind in Anhang I (siehe Tabelle 88 und Tabelle 89) dargestellt.

Insgesamt lässt sich auf Grundlage der gruppenspezifischen Analyse der Kausalstruktur feststellen, dass das untersuchte Kausalmodell auch in der Gruppe der Sekundarstufenlehrkräfte mit Blick auf VERA8 gültig ist. Die Schätzung des gemeinsamen Strukturmodells unterstreicht des Weiteren erneut die Bedeutung der wahrgenommen Nützlichkeit und insbesondere deren direkten Einfluss auf die Nutzungsintention, der bereits bei der Eingruppentestung in der VERA3-Stichprobe deutlich wurde. Diese Erkenntnis wird im folgenden Diskussionsteil, neben den anderen Ergebnissen, noch einmal ausführlich aufgegriffen.

6. Diskussion

In diesem Kapitel werden die zentralen Befunde der theoretischen und empirischen Untersuchung dieser Arbeit vor dem Hintergrund der forschungsleitenden Fragestellungen zusammengefasst und abschließend kritisch beleuchtet. Hierbei erfolgt ein erneuter Rückgriff auf den theoretischen Rahmen der Arbeit sowie eine Einordnung in den in Kapitel 2.2.2 skizzierten Forschungsstand zu Vergleichsarbeiten. Aus den Erkenntnissen werden zudem Implikationen für die Praxis der Vergleichsarbeiten abgeleitet. Des Weiteren werden die methodischen Limitationen der Arbeit diskutiert und Perspektiven für die weitere Forschung aufgezeigt.

6.1. Zusammenfassende Diskussion zentraler Befunde

Übergeordnetes Ziel dieser Arbeit war es, im Zuge einer Erweiterung der Akzeptanzforschung die Wahrnehmung und insbesondere die nur sehr dürftige Nutzung von Vergleichsarbeiten durch Lehrkräfte zu erklären und mögliche Ursachen zu identifizieren und theoretisch einzuordnen. Ein besonderes Anliegen war dabei, im Kontext von VERA eine begriffliche Klarheit um das Konstrukt Akzeptanz zu schaffen und dieses empirisch zu untersuchen. Konkret umfassten die Ziele die Entwicklung einer theorie- und empiriegeleiteten Definition von Akzeptanz (Ziel 1), die Konzeption eines empirisch überprüfbaren Akzeptanzmodells (Ziel 2) und daran anschließend die Prüfung und Validierung dieses Modells (Ziel 3). Darüber hinaus beschäftigte sich die Arbeit mit der Frage nach Unterschieden in der Akzeptanz bei Lehrkräften der verschiedenen Schularten, konkret dem Vergleich von VERA3 und VERA8 und der damit einhergehenden Untersuchung einer allgemeinen Gültigkeit des aufgestellten Akzeptanzmodells im Zusammenhang mit VERA (Ziel 4). Die folgenden Unterkapitel fassen die zentralen Erkenntnisse der vorliegenden Arbeit vor dem Hintergrund dieser Forschungsziele zusammen.

6.1.1. Theoriegeleitete Definition von Akzeptanz und Konzeption eines Forschungsmodells im Kontext von VERA (Ziel 1 und 2)

Bei der Auseinandersetzung mit der Forschungsliteratur zur Sicht von Lehrkräften in Bezug auf Vergleichsarbeiten stellte sich schnell heraus, dass sich zwar viele Arbeiten mit der

Akzeptanz von Vergleichsarbeiten beschäftigen, es jedoch an einer theoretischen Grundlage und einer klaren Definition mangelt. Zwar wird der Begriff Akzeptanz i. d. R. nicht explizit definiert, häufig jedoch in Verbindung mit Einstellung, Nützlichkeit bzw. Nutzen und auch Nutzung gebraucht. Auch wenn sich das Akzeptanzverständnis in nahezu allen Forschungsarbeiten nur aus der Operationalisierung der Konstrukte oder der Interpretation der Ergebnisse ableiten lässt, wird deutlich, dass Akzeptanz entsprechend affektive (einstellungsbezogene) (siehe bspw. Ditton et al., 2002; Maier, 2008a, 2008b, 2009c), kognitive (häufig in Verbindung mit einer Nutzenbewertung) (siehe bspw. Wurster & Richter, 2016; Wurster, Bach et al., 2016) und auch behaviorale Elemente umfassen kann und das Verhalten von Lehrkräften im Hinblick auf VERA beeinflusst (siehe bspw. Groß Ophoff et al., 2019; Helmke & Hosenfeld, 2005; Koch, 2011; Vogel et al., 2016). Diese Aspekte sind auch in verschiedenen Einstellungstheorien wie dem Drei-Komponenten-Modell (Rosenberg & Hovland, 1969), der TRA (Theory of Reasoned Action) nach Fishbein und Ajzen (1975), der TPB (Theory of Planned Behavior) nach Ajzen (1985, 1991) und dem TAM (Technology Acceptance Model) nach F. D. Davis (1986) und F. D. Davis et al. (1989) zu finden. Daher wurde auf diese Einstellungstheorien (siehe Kapitel 2.1.2), bzw. speziell das TAM, zurückgegriffen, um zunächst eine Definition von Akzeptanz und im nächsten Schritt, in Verbindung mit den in der Literatur zu Vergleichsarbeiten identifizierten Wahrnehmungsfaktoren (siehe Kapitel 2.2.2), ein empirisch überprüfbares Akzeptanzmodell zu entwickeln.

TRA und TPB nehmen eine Neuordnung der affektiven, kognitiven und behavioralen Komponenten von Einstellung vor, indem sie Verhalten als ein einer Einstellung nachgelagertes Konstrukt begreifen und des Weiteren hinsichtlich der Verhaltenskomponenten zwischen Verhaltensintention und daraus resultierendem Verhalten unterscheiden (siehe bspw. Eagly & Chaiken, 2005). Diese Theorien bilden zwar einen guten Ausgangspunkt zur Erklärung der Struktur der verschiedenen Aspekte von Einstellungen bzw. zur Erklärung von Verhalten, die sich auch in Grundzügen in der Literatur zur Wahrnehmung und Bewertung von VERA wiederfinden, der Begriff Akzeptanz findet jedoch in diesen Modellen noch keinen Platz. Erst mit der Weiterentwicklung der Modelle durch das TAM taucht im Kontext der Einstellungs- bzw. Verhaltenstheorien zum ersten Mal der Begriff Akzeptanz auf. Das TAM folgt der Logik, dass zunächst der Kontakt mit einem Einstellungsobjekt eine kognitive Reaktion auslöst, die zur Entstehung einer affektiven Reaktion in Form einer Einstellung führt, welche eine Verhaltensreaktion auslöst (F. D. Davis, 1986; F. D. Davis et al., 1989).

Basierend auf dieser modellinhärenten Prozesslogik wurde eine Definition von Akzeptanz aufgestellt und die erste Forschungsfrage dieser Arbeit nach einem theoriebasierten Begriffsverständnis und konkreten Definition beantwortet: Akzeptanz bildet demnach nicht nur ein einfaches Konstrukt ab, sondern einen gesamten Wahrnehmungs- und Verarbeitungsprozess, welcher kognitive, affektive und behaviorale Komponenten umfasst und versteht sich als die *Bildung einer positiven (affektiven) Einstellung gegenüber einem Akzeptanzobjekt auf Basis kognitiver Prozesse, die in einer entsprechenden positiv gerichteten Verhaltensabsicht bzw. einem Verhalten resultiert.*

Diese Definition von Akzeptanz ist positiv konnotiert und spiegelt somit einen (positiven) Wahrnehmungs- und Bewertungsprozess wider, der immer auch einen positiven Verhaltensaspekt umfasst. Ohne eine entsprechende Nutzung oder zumindest eine konstatierte Nutzungsintention kann demnach nicht von Akzeptanz gesprochen werden.

Mit dem Ziel dieser Arbeit, die begriffliche Konfusion rund um die Untersuchung von Akzeptanz im Kontext von Vergleichsarbeiten aufzulösen, wurde im nächsten Schritt ein Modell zur Konzeptualisierung von Akzeptanz im Kontext von VERA entwickelt, welches die in der Literatur identifizierten Konstrukte (siehe Kapitel 2.2) den Modellannahmen des TAM folgend ordnet. Auf Grundlage der aus dem TAM abgeleiteten Akzeptanzdefinition und den Erkenntnissen aus der Aufbereitung des Forschungsstandes zur Wahrnehmung und Akzeptanz von Vergleichsarbeiten bei Lehrkräften wurde daher ein Forschungsmodell zur Erklärung der verschiedenen Akzeptanzfaktoren von VERA aufgestellt. Aufgrund dessen, dass sich aus der VERA-Literatur alleine keine eindeutigen Wirkungsbeziehungen zwischen den einzelnen Konstrukten identifizieren ließen, wurden die Zusammenhänge zwischen den Einflussgrößen gemäß der Logik des TAM postuliert (siehe Abbildung 23, S. 232). Hierbei fanden jedoch zunächst nicht alle identifizierten Faktoren in dem letztendlich aufgestellten Forschungsmodell einen Platz, da es zunächst das Ziel war, ein erstes Grundmodell in Anlehnung an das ursprüngliche TAM zu entwickeln und zu validieren. Mögliche Modellerweiterungen werden entsprechend in Kapitel 6.4 diskutiert.

Die Definition des Akzeptanzbegriffs und die Entwicklung dieses Forschungsmodells schließen in der Forschungsliteratur die Lücke einer fehlenden Definition und Konzeptualisierung des bisher im Kontext von Vergleichsarbeiten nur diffus verwendeten Konstrukts Akzeptanz.

Die theoriebasierte Definition von Akzeptanz auf Grundlage des TAM stellt einen wichtigen ersten Schritt dar, Klarheit und Stringenz in dem bisher wenig strukturiert beforschten Feld, der Akzeptanzforschung im Bereich von Vergleichsarbeiten, zu schaffen. Ein kongruentes Begriffsverständnis schafft hierbei eine gemeinsame Basis, auf deren Grundlage Anschlussuntersuchungen angestoßen und wissenschaftlicher Fortschritt erzielt werden können. Das Modell ermöglicht eine theoriegeleitete empirische Überprüfung der Beziehungen verschiedener Wahrnehmungs- bzw. Einstellungskonstrukte, deren Zusammenhänge bisher in der Literatur nicht eindeutig geklärt werden konnten. In der Literatur bestehen demnach häufig Überschneidungen und Durchmischungen von Begriffen und Operationalisierungen, indem bspw. gleich operationalisierte Konstrukte unterschiedlich benannt oder gleich benannte Konstrukte verschieden operationalisiert werden (siehe Kapitel 2.2). Das Modell stellt somit einen ersten Ansatz zur strukturierten Analyse des bisher diffus verwendeten Konstrukts Akzeptanz dar.

6.1.2. Empirische Validierung des Akzeptanzmodells (Ziel 3)

Die statistische Überprüfung des aufgestellten Modells stellte das dritte Forschungsziel der vorliegenden Arbeit dar. Hierbei wurde zunächst das ursprünglich aufgestellte Modell (siehe Abbildung 23) anhand einer Stichprobe aus VERA3-Lehrkräften (Modell $1_{(V3\text{-}18)}$ bzw. Modell $1_{dP(V3\text{-}18)}$) vor dem Hintergrund der postulierten Hypothesen analysiert (siehe Kapitel 5.1). Datengrundlage der ersten Modellprüfung bildeten Survey-Daten aus Befragungen von VERA3-Lehrkräften des Jahres 2018. Dieses Modell erwies sich zwar durchaus als passend zu den Daten, jedoch ergaben sich Hinweise für eine datengeleitete Anpassung des ursprünglichen Modells. Während das ursprüngliche Modell 1 somit verworfen wurde, bewährte sich das angepasste Modell 2 sowohl in der Ursprungsstichprobe (Modell $2_{(V3\text{-}18)}$ bzw. Modell $2_{dP(V3\text{-}18)}$) (siehe Kapitel 5.2) als auch bei der Validierung mit einer weiteren unabhängigen Stichprobe der VERA3-Lehrkräftebefragung des Jahres 2019 (Modell $2_{(V3\text{-}19)}$ bzw. Modell $2_{dP(V3\text{-}19)}$) (siehe Kapitel 5.3). Da sich dieses Modell 2 als besser zu den Daten passend und letztendlich inhaltlich plausibler herausstellte, wurde dieses Modell für die weiteren Analysen beibehalten und wird im Folgenden diskutiert. Vor der Prüfung des Kausalmodells wurde die Passung der einzelnen Konstrukte anhand von CFAs überprüft. Die Ergebnisse dieser Schätzungen werden zunächst im folgenden Unterkapitel mit Blick auf aufgetretene Probleme und Besonderheiten diskutiert.

Skalenvalidierung

Wie die Ausführungen in Kapitel 5.1.2 und 5.3 bzw. Anhang E verdeutlichen, erweist sich die Passung nahezu aller entwickelten Skalen auf Basis der VERA3-Daten der Jahre 2018 und 2019 als gut bzw. sehr gut. Die Konstrukte Nutzungsintention, Einstellung, Aufwand-Nutzen sowie Nützlichkeit weisen durchweg zufriedenstellende Werte der Itemkorrelationen, Trennschärfe, internen Konsistenz und DEV, sowie gute bzw. mindestens akzeptable Fitstatistiken bzw. Parameterschätzungen basierend auf den durchgeführten konfirmatorischen Faktorenanalysen auf.[9] Lediglich die Modellierung des Konstrukts zeitliche Belastung gestaltete sich nicht unproblematisch und soll daher an dieser Stelle noch einmal detaillierter beleuchtet werden: Zunächst konnte das Konstrukt aufgrund der geringen Anzahl an Indikatorvariablen nicht in einem separaten Messmodell geschätzt werden, da dieses ohne weitere Restriktionen nur gerade identifiziert wäre. Daher erfolgte die Schätzung in einem gemeinsamen Messmodell mit dem Konstrukt Nützlichkeit. Der Gesamtmodellfit weist zwar durchweg gute Kennwerte auf, der Blick auf die Parameterschätzungen des Konstrukts zeitliche Belastung offenbart jedoch die Schwächen der Modellierung. Bereits die manifesten Itemkorrelationen von .19 bis .43 weisen darauf hin, dass der geringe Zusammenhang zwischen den Items ein Problem darstellen könnte, ebenso kennzeichnet der DEV von .32 eine nicht optimale Repräsentation der Skala durch die Indikatorvariablen.

Hinsichtlich der Faktorladungen weist lediglich Item ZB2 (Zeitaufwand der Durchführung) mit $\lambda^s \geq .803$ eine als hoch zu wertende Faktorladung auf, während die Items ZB1 (Zeitaufwand der Vorbereitung) und ZB3 (Zeitaufwand der Auswertung) lediglich die Minimalanforderungen von $\lambda^s \geq .40$ (Berning, 2019) bzw. $\lambda^s \geq .50$ (Urban & Mayerl, 2014) erfüllen. Ebenso erzielt nur Item ZB2 mit 65 % eine zufriedenstellende Varianzaufklärung, während die Items ZB1 und ZB3 mit 16 % bzw. 29 % deutlich unter den angestrebten 50 % liegen. Insgesamt wird das Konstrukt somit am besten durch Item ZB2 repräsentiert, welches den Zeitaufwand der Testdurchführung erfragt, während die Items ZB1 und ZB3 das Konstrukt nicht optimal repräsentieren. Auch wenn dieses Ergebnis, wie bereits ausführlich in Kapitel 5.1.2 dargelegt, auf eine mögliche reflektive Fehlmodellierung hinweist, wobei die

[9] Die im Folgenden berichteten Zahlen beziehen sich auf die in Kapitel 5.1.2. berichtete Skalenanalyse mit den VERA3 Daten des Jahres 2018, die Ergebnisse der Stichprobe des Jahres 2019 unterscheiden sich jedoch nur geringfügig (siehe Anhang E).

Beweislage bei weitem nicht eindeutig ist, soll an dieser Stelle noch einmal betont werden, dass die größte Problematik, die durch eine fehlerhafte reflektive Modellierung entstehen könnte, für dieses Modell nicht relevant ist: Diese Problematik betrifft die mögliche Gefährdung der Inhaltsvalidität durch das Entfernen schlecht fittender Items. Da jedoch keines der drei Items des Konstrukts zeitliche Belastung entfernt wurde, entfällt die Relevanz.

Dennoch lohnt sich noch ein Blick auf die deskriptiven Itemstatistiken der Indikatorvariablen: Diese verdeutlichen, dass der Vorbereitungsaufwand insgesamt am niedrigsten bewertet wird ($M = 2.75$, $SD = 0.95$), während die befragten Lehrkräfte den Zeitaufwand der Auswertung am höchsten einschätzen ($M = 3.42$, $SD = 0.83$), den Durchführungsaufwand als durchschnittlich und somit angemessen ($M = 3.02$, $SD = 0.73$). Dass der Aufwand der Auswertung am höchsten und dabei auch im Schnitt als eher zu hoch empfunden wird, entspricht den Erkenntnissen früherer Untersuchungen, die besonders die mit der Auswertung und Datenübermittlung einhergehende zeitliche Belastung bemängeln (siehe bspw. Diemer, 2013). Die eher niedrige Bewertung des Vorbereitungsaufwands kann dadurch erklärt werden, dass die Lehrkräfte i. d. R. fertig gedruckte und vorbereitete Testmaterialien erhalten, weil diese in den meisten Bundesländern zentral gedruckt und an die Schulen geliefert werden oder, falls dies nicht der Fall ist, häufig vom Sekretariat vorbereitet werden. Infolgedessen fällt der reale Vorbereitungsaufwand in der Praxis eher gering aus. Die neutrale Bewertung des Durchführungsaufwandes erklärt sich womöglich dadurch, dass die Unterrichtsstunde ohnedies stattfinden müsste und durch die Testdurchführung kein unmittelbarer Mehraufwand entsteht.

Insgesamt zeichnet sich mit Blick auf die einzelnen Aspekte ein recht heterogenes Bild der Bewertung des Zeitaufwandes der verschiedenen Arbeitsschritte, wodurch sich ergänzende inhaltliche Hinweise zur Erklärung der nicht optimalen Passung des Konstrukts zeitliche Belastung ergeben. Abgesehen von dem Konstrukt zeitliche Belastung sprechen die Ergebnisse für eine insgesamt adäquate Operationalisierung der Modellkonstrukte. Dies bestätigt sich auch in der Validierung der Konstrukte des angepassten Modells 2 anhand der VERA3-Evaluationsdaten des Jahres 2019 (siehe Kapitel 5.3 und Anhang E).

Kausalanalyse

Auch wenn sich in der Analyse verschiedener Modelle Modell 2 gegenüber dem ursprünglich aufgestellten Modell 1 als zu bevorzugen erwies, werden beide Modelle mit Fokus auf die

daraus gewonnenen zentralen Erkenntnisse in diesem Abschnitt noch einmal vergleichend gegenübergestellt. Abbildung 23 visualisiert das ursprüngliche Modell 1 inklusive des aufgestellten Hypothesensystems, Abbildung 24 das finale Modell 2_{dp} im Vergleich der beiden VERA3-Stichproben 2018 und 2019.

Die farbliche Hinterlegung der postulierten Vorzeichen in Abbildung 23 kennzeichnet dabei, ob ein Pfad die vermutete Richtung bzw. Stärke aufwies. Grün markiert die empirische Bestätigung der hypothetisierten Zusammenhänge (ausgedrückt als Vorzeichen), rot die empirische Zurückweisung der hypothetisierten Zusammenhänge. Im Detail wird jedoch an dieser Stelle nicht mehr auf die Hypothesen eingegangen, lediglich Auffälligkeiten werden hervorgehoben. Wie in Kapitel 5.1.3 ausführlich beschrieben, konnte ein Großteil der aufgestellten Hypothesen anhand beider Varianten von Modell 1 bestätigt werden. Für eine übersichtliche detaillierte Darstellung wird an dieser Stelle erneut auf Tabelle 37 auf S. 190 verwiesen. Alle ge-samten indirekten Effekte sowie die Gesamteffekte entsprechen hinsichtlich ihrer Richtung und ihrer Signifikanz den Erwartungen. Auffällig erweisen sich lediglich die in Abbildung 23 rot hinterlegten Pfade, welche die Hypothesen H3d und H6d repräsentieren. Diese verdeutlichen, dass in Modell 1 kaum ein direkter Effekt von wahrgenommener Nützlichkeit und zeitlicher Belastung auf die Einstellung zu finden ist, sondern beide Konstrukte nahezu ausschließlich indirekt über das Aufwand-Nutzen-Konstrukt wirken. Da diese nicht erwartungskonformen Effekte Hinweise für eine Modifizierung des Modells in Richtung Modell 2 gaben, werden diese hier noch einmal näher in den Blick genommen.

Bei der Betrachtung der Wirkung der zeitlichen Belastung auf die Einstellung der Lehrkräfte zeigt sich zunächst nicht wie vermutet ein direkter negativer Effekt, sondern ein schwach positiver Effekt. Eine inhaltliche Erklärung für diesen positiven direkten Effekt und die daraus resultierende Ablehnung von Hypothese H3d könnte in dem Bestreben von Menschen liegen, kognitive Dissonanzen zu reduzieren (Festinger, 1957). Wurde erst einmal ein zeitlicher Aufwand investiert, um ein als wenig nützlich erachtetes Instrument einzusetzen, führt dies ggf. zur Anpassung der Einstellung und der Selbstüberzeugung, dass der Aufwand nicht umsonst war.

Der Gesamteffekt der zeitlichen Belastung auf die Einstellung, mediiert durch Nützlichkeit und Aufwand-Nutzen, erweist sich als erwartungsgemäß negativ. Dies spricht dafür, dass der

positive Effekt durchaus auch modellinhärente Ursachen haben könnte, die bspw. durch den zuvor beschriebenen starken Zusammenhang zwischen den Konstrukten Aufwand-Nutzen und Nützlichkeit sowie insbesondere zwischen Aufwand-Nutzen und Einstellung verursacht sein könnten. Dieser Zusammenhang schlägt sich auch im Strukturmodell nieder und beeinflusst dadurch die Höhe der übrigen Modellpfade. Ein zu stark geschätzter Effekt könnte hierbei bei anderen Pfaden einen korrigierenden Einfluss in eine andere Richtung hervorrufen.

Auch im Hinblick auf die Wirkung der wahrgenommenen Nützlichkeit erfolgt der Großteil des Einflusses des Konstrukts auf die Einstellung, wie bereits die hohe Faktorkorrelation vermuten lässt, über den indirekten Pfad der Wirkung der Nützlichkeit auf die Aufwand-Nutzen-Bewertung. Der direkte Einfluss der Nützlichkeit auf die Einstellung ist in diesem Modell entgegen den Erwartungen quasi nicht existent. Der auch im TAM postulierte Einfluss der wahrgenommenen Nützlichkeit auf die Einstellung bestätigt sich somit auch in diesem Modell, jedoch nur indirekt gemittelt über die Aufwand-Nutzen-Abwägung.

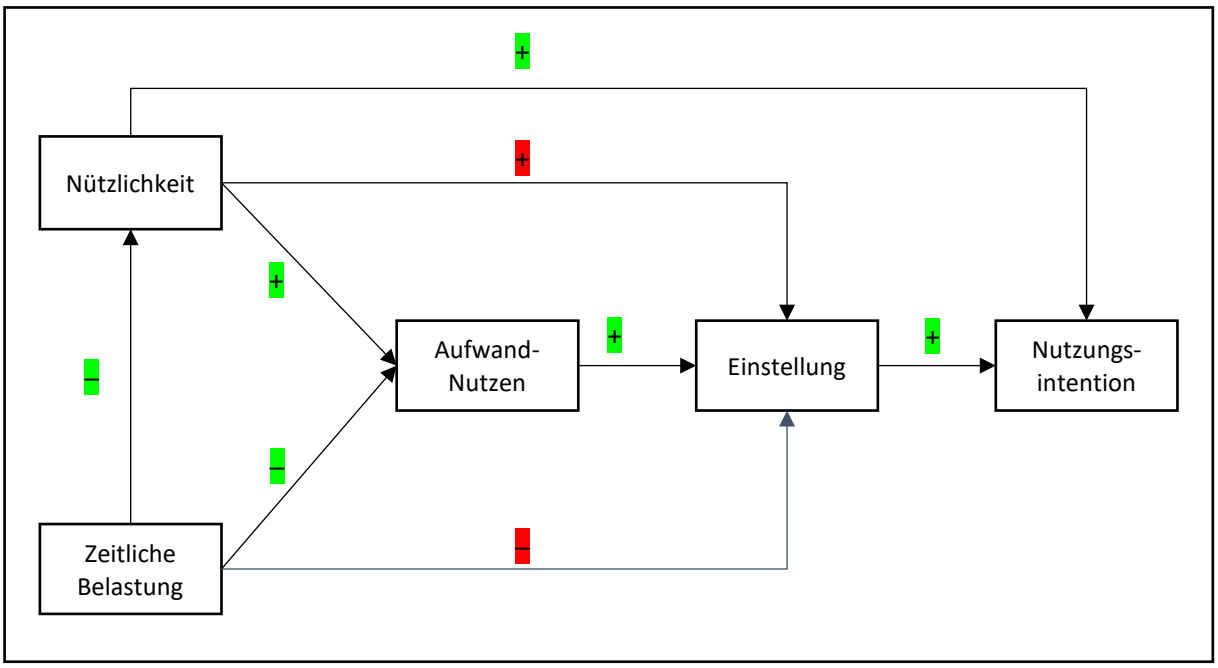
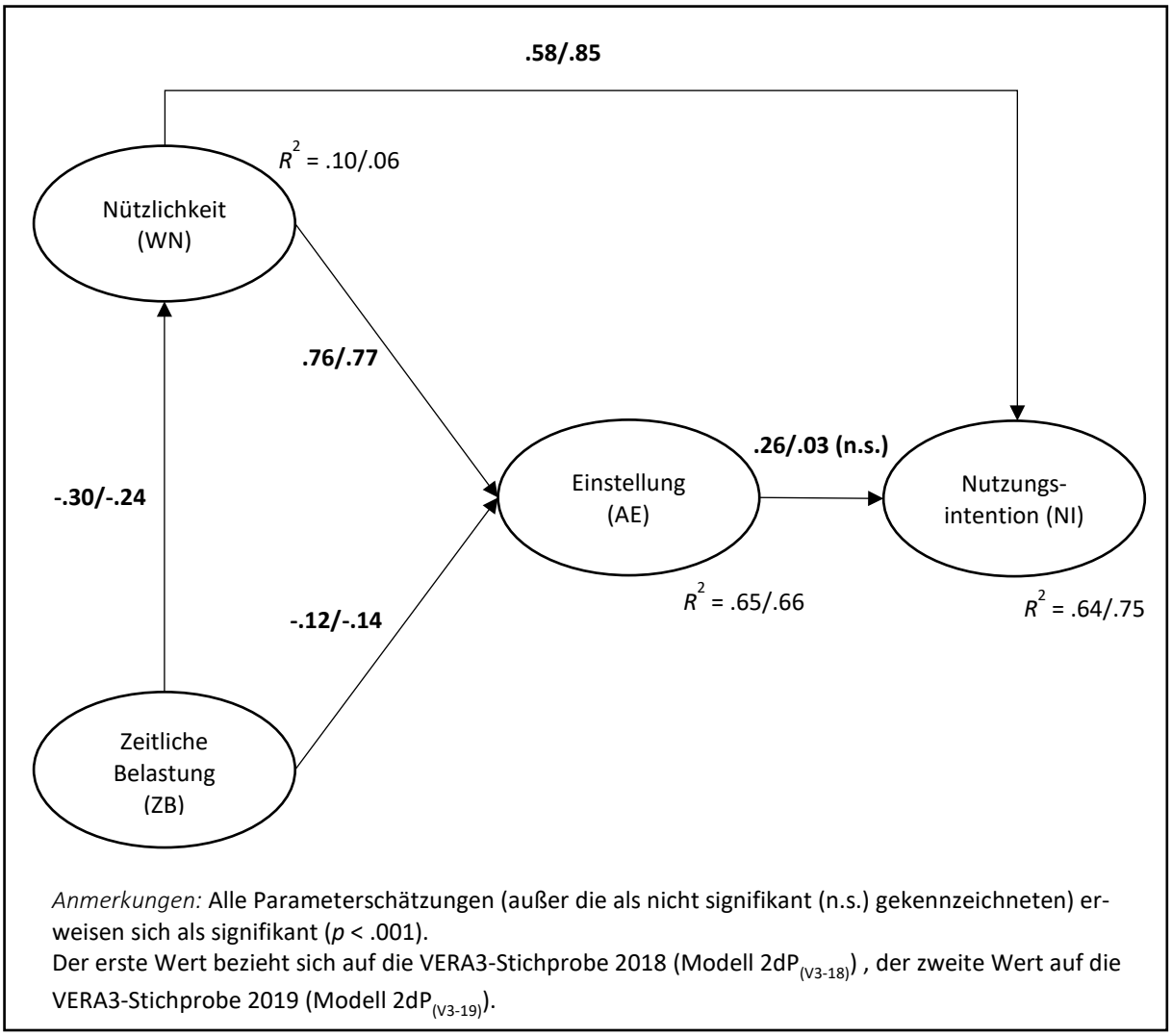


Abbildung 23: Forschungsmodell (Modell 1) inklusive postulierter Effekte

Zwar erweisen sich die direkten Pfade sowohl von Nützlichkeit als auch zeitlicher Belastung auf die Einstellung als nicht signifikant bzw. weisen das falsche Vorzeichen auf, dennoch wirken beide Konstrukte in der postulierten Richtung zumindest indirekt über das Konstrukt Aufwand-Nutzen. Da sich dieses Konstrukt inhaltlich direkt aus diesen beiden vorgelagerten

Konstrukten zeitliche Belastung und Nützlichkeit speist, ist dieses Ergebnis nicht weiter verwunderlich. Es stellte sich jedoch die Frage, ob die Spezifizierung einer Aufwand-Nutzen-Abwägung als separates Konstrukt überhaupt notwendig ist oder ob dem nicht bereits durch das Zusammenwirken der Konstrukte zeitliche Belastung und Nützlichkeit im Modell Rechnung getragen wurde. Vor dem Hintergrund der dazukommenden statistischen Herausforderungen durch die starken Interfaktorkorrelationen zwischen dem Aufwand-Nutzen-Konstrukt und der Einstellung sowie der Nützlichkeit wurde das Modell in einem weiteren Schritt überarbeitet und das Konstrukt Aufwand-Nutzen entfernt (siehe folgender Abschnitt und Kapitel 5.2).

Basierend auf den Ergebnissen der Modellschätzung, zusammen mit den inhaltlichen Überlegungen, erfolgte durch die Entfernung des separaten Aufwand-Nutzen-Konstrukts die Anpassung des ursprünglichen Modells hin zu dem final bevorzugten Modell 2 (siehe Abbildung 24). Der vergleichende Blick auf die grafische Darstellung der Modelle (siehe Abbildung 23, Modell 1 und Abbildung 24, Modell 2) verdeutlicht, dass das angepasste Modell 2 im Hinblick auf die Wirkungszusammenhänge zwischen den Konstrukten stärker dem Aufbau des ursprünglichen TAM entspricht und die postulierten direkten Pfade von zeitlicher Belastung und Nützlichkeit auf die Einstellung, welche in Modell 1 (siehe rot markierte Effekte) nicht bestätigt werden konnten, sich in diesem Modell auch empirisch bewähren. Das angepasste Modell 2 passt besser zu den Daten als das Ursprungsmodell und bewährt sich auch in der Validierung mit einer unabhängigen Stichprobe (siehe Kapitel 5.3 Modellvalidierung mit VERA3 2019). Auch in diesem Modell wurde durch die Spezifizierung des direkten Pfades zwischen Nützlichkeit und Nutzungsintention eine nicht unwesentliche Modellverbesserung erzielt (siehe Fitstatistiken Kapitel 5.3, Tabelle 41, S. 199).



Anmerkungen: Alle Parameterschätzungen (außer die als nicht signifikant (n.s.) gekennzeichneten) erweisen sich als signifikant ($p < .001$).
Der erste Wert bezieht sich auf die VERA3-Stichprobe 2018 (Modell $2dP_{(V3\text{-}18)}$) , der zweite Wert auf die VERA3-Stichprobe 2019 (Modell $2dP_{(V3\text{-}19)}$).

Abbildung 24: Parameterschätzungen der Modelle $2_{dP(V3\text{-}18)}$ und $2_{dP(V3\text{-}19)}$ (vollständig standardisierte Lösung).

Der Blick auf die Modellpfade des finalen Modells 2_{dP}, das in Abbildung 24 im Vergleich der beiden VERA3-Stichproben 2018 und 2019 dargestellt ist, legt mit dem TAM vergleichbare Kausalbeziehungen offen. Erwartungsgemäß wirken Nützlichkeit und zeitliche Belastung auf die Einstellung. Die zeitliche Belastung wirkt dabei zum einen mit einem eher kleinen Effekt direkt negativ auf die Einstellung und zum anderen indirekt über ihren etwas stärkeren Einfluss auf die wahrgenommene Nützlichkeit. Diese hat wiederum einen relativ starken direkten Effekt auf die Einstellung. Die Höhe der Effekte ist zwischen den beiden Stichproben in etwa vergleichbar. Unterschiede zwischen den beiden Erhebungsjahren zeigen sich durch die Spezifizierung des direkten Pfades zwischen Nützlichkeit und Nutzungsintention. Der Einfluss der Einstellung auf die Nutzungsintention nimmt in beiden Gruppen ab, wird jedoch in der Validierungsstichprobe (VERA3 2019) insignifikant, während der direkte Effekt der

Nützlichkeit auf die Nutzungsintention sehr stark ausfällt (β = .85). In der VERA3-Stichprobe 2018 verliert der Effekt der Einstellung dagegen nicht vollständig an Bedeutung (β = .26), während die Wirkung der Nützlichkeit auf die Nutzungsintention mit β = .58 etwas schwächer ausfällt. Die Tendenzen der Effekte im Sinne von Wirkungsrichtung und Stärke sind somit auch zwischen verschiedenen Stichproben vergleichbar. Die ermittelten Unterschiede zwischen den betrachteten Jahren bedürfen ggf. noch weiterer Untersuchungen. Möglicherweise spielen auch noch weitere Faktoren, die in diesem Modell zunächst keine Berücksichtigung finden, eine Rolle bei der Bewertung der Vergleichsarbeiten durch Lehrkräfte. Mögliche Modellerweiterungen werden im Forschungsausblick in Kapitel 6.4 noch detaillierter in den Blick genommen.

Insgesamt lässt sich jedoch resümieren, dass Modell 2 gegenüber dem theoriegeleiteten Ursprungsmodell zu bevorzugen ist, da es zum einen eine höhere Gesamtmodellpassung aufweist und sich zum anderen in verschiedenen Stichproben bewährt. Sowohl der Modellfit als auch die Modellpfade erweisen sich als weitestgehend stabil zwischen den Jahren. Diese Ergebnisse liefern erste Hinweise für eine allgemeine Gültigkeit des Modells und bilden die Voraussetzung für die weitere Anwendung in zukünftigen Untersuchungen sowie u. a. die in Kapitel 5.4 beschriebene Gruppenanalyse. Es wird deutlich, dass auf die explizite Modellierung eines Aufwand-Nutzen-Konstrukts verzichtet werden kann, auch wenn die Literatur zu Vergleichsarbeiten die Relevanz dieses Konstrukts bei der Wahrnehmung und Bewertung von VERA durch Lehrkräfte nahelegt. In der empirischen Überprüfung erweist sich jedoch die Abbildung dieses Phänomens durch das Zusammenwirken der Konstrukte zeitliche Belastung und Nützlichkeit (Modell 2) als dem komplexeren Modell mit separater Konstruktmodellierung (Modell 1) überlegen.

6.1.3. Analyse von Unterschieden in der Akzeptanz bei Lehrkräften verschiedener Schularten (Ziel 4)

Ziel des letzten Parts des Empirieteils war es, mögliche Unterschiede in der Wahrnehmung von Vergleichsarbeiten bei Lehrkräften verschiedener Schulformen aufzudecken, also konkret Unterschiede zwischen Grundschullehrkräften (VERA3) und Lehrkräften weiterführender Schulen (VERA8) zu analysieren. Da der Forschungsstand zeigte, dass es bisher nur wenig empirische Erkenntnisse zu dieser Fragestellung gibt, hatte dieser letzte Teil einen eher

explorativen Charakter. Es stand lediglich die Vermutung im Raum, dass Grundschullehrkräfte Vergleichsarbeiten grundsätzlich tendenziell etwas positiver gegenüberstehen als Lehrkräfte weiterführender Schulen, VERA3 somit größeren Anklang unter Lehrkräften findet als VERA8. Datengrundlage der Untersuchung bildete eine Teilstichprobe der VERA3-Stichprobe aus 2018 und die Evaluationsbefragung mit VERA8-Lehrkräften im Nachgang der VERA-Testung desselben Jahres (siehe auch Kapitel 4.2.5). Zur Untersuchung der Stichprobenunterschiede wurde eine latente Gruppenanalyse durchgeführt. Hierfür wurden zunächst die Messmodelle auf Invarianz untersucht, bevor auf Basis eines gemeinsamen Messmodells die latenten Mittelwerte und Pfadkoeffizienten anhand des angepassten Strukturmodells 2 untersucht werden konnten. Eine Zusammenfassung der Ergebnisse unter Rückgriff auf die anfängliche Vermutung einer positiveren Bewertung von VERA3 im Vergleich zu VERA8 liefern die folgenden Abschnitte.

Fragestellung 1: Skalenvalidierung

Nach einer Item- und Reliabilitätsanalyse in Kapitel 5.4.1, erfolgte in Kapitel 5.4.2 eine Skalenvalidierung mittels konstruktweiser Messinvarianzanalyse mit dem Ziel, zunächst die für VERA3-Lehrkräfte gültigen Messmodelle anhand einer Stichprobe der VERA8-Lehrkräfte zu validieren. Für die Konstrukte Nutzungsintention und Einstellung konnte jeweils das Vorliegen strikter Invarianz nachgewiesen werden. Auch die separate Schätzung des Konstrukts Nützlichkeit führte unter den Bedingungen strikter Invarianz zu einem guten Modellfit. Die Schätzung des Konstrukts zeitliche Belastung bzw. die gemeinsame Schätzung mit dem Konstrukt Nützlichkeit erwies sich dagegen wie schon bei der Schätzung der Messmodelle mit den VERA3-Stichproben (siehe Kapitel 5.1.2 und 5.3 bzw. Anhang E) als problematisch. Bereits die Modellierung skalarer Invarianz im gemeinsamen Messmodell scheiterte an einem unzureichenden Modellfit. Jedoch konnte durch die Freisetzung des Intercepts von Item ZB3 (Zeitaufwand der Auswertung) zunächst partielle skalare Invarianz und im nächsten Prüfschritt mit freier Schätzung der Residualvarianz desselben Items auch partielle strikte Invarianz nachgewiesen werden.

Formell wurde somit für alle Konstrukte durch den Nachweis (partiell) skalarer oder sogar strikter Invarianz die Voraussetzung zur Analyse latenter Mittelwertdifferenzen zwischen den betrachteten Gruppen, die im nächsten Schritt durchgeführt wurde, erfüllt. Trotz Vorliegen der formalen Voraussetzungen wurde aufgrund inhaltlicher und statistischer Überlegungen im

Weiteren darauf verzichtet, für das Konstrukt zeitliche Belastung latente Mittelwertunterschiede zu analysieren. Daher wurden in der abschließenden Schätzung eines gemeinsamen Messmodells alle herausgearbeiteten Restriktionen des Konstrukts zeitliche Belastung berücksichtigt. Ein finales Messmodell mit partieller strikter Invarianz (Modell $\text{GroupCFA}_{\text{strInv-partiell b}}$) wurde geschätzt. In diesem Modell wurden die Intercepts des latenten Konstrukts zeitliche Belastung in beiden Gruppen auf null fixiert und somit für dieses Konstrukt keine Mittelwertdifferenzen geschätzt. Dieses Modell wies insgesamt gute Fitstatistiken auf ($\chi_r^2(303) = 962.63$, $p < .001$, $\chi^2/df = 3.18$, $\text{CFI}_r = .976$, $\text{TLI}_r = .976$, $\text{RMSEA}_r = .043$, SRMR = .043) und wurde aufgrund inhaltlicher Plausibilitätsüberlegungen gegenüber einem Modell, das für alle Konstrukte inklusive zeitliche Belastung latente Mittelwerte schätzt (Modell $\text{GroupCFA}_{\text{strInv-partiell a}}$), bevorzugt und für die weiteren Analysen (siehe Kapitel 5.4.3 und 5.4.4) genutzt.

Insgesamt konnte die Prüfung der verschiedenen Stufen der Messinvarianz bis hin zum Nachweis strikter oder zumindest partiell strikter Invarianz zeigen, dass die untersuchten Konstrukte in beiden analysierten Gruppen dieselben Sachverhalte messen. Die gewählte Operationalisierung scheint somit geeignet, die Wahrnehmung, sowohl von Grundschullehrkräften im Hinblick auf VERA3, als auch von Lehrkräften weiterführender Schulen hinsichtlich VERA8, adäquat zu erfassen. Die Ergebnisse der Untersuchung liefern somit erste Hinweise auf die Allgemeingültigkeit der genutzten Skalen.

Fragestellung 2: Mittelwertdifferenzen

Das Vorliegen von Messinvarianz, das heißt der Nachweis der Gültigkeit der Messmodelle in verschiedenen Stichproben, ermöglichte im nächsten Schritt die Analyse von Unterschieden in der Wahrnehmung der verschiedenen Gruppen von Lehrkräften auf Ebene der latenten Konstrukte. Hierfür wurden auf Basis von Modell $\text{GroupCFA}_{\text{strInv-partiell a}}$ jeweils die latenten Mittelwerte bzw. die Intercepts der latenten Konstrukte in der VERA3-Gruppe als Referenzgruppe auf null fixiert. Ausnahme bildete, wie dargelegt, das Konstrukt zeitliche Belastung, für das keine latenten Mittelwerte geschätzt wurden, indem die Intercepts beider Gruppen auf null festgesetzt wurden. Eine ausführliche Darstellung der Ergebnisse findet sich in Kapitel 5.4.3.

Die Analyse der latenten Mittelwerte der Konstrukte Nützlichkeit, Einstellung und Nutzungsintention ergibt für alle Konstrukte eine signifikante Mittelwertdifferenz mit einem jeweils negativen Vorzeichen. Die latenten Mittelwerte fallen somit für alle drei Konstrukte in der Gruppe der VERA8-Lehrkräfte signifikant niedriger aus als in der Gruppe der VERA3-Lehrkräfte. Die Analyse der latenten Skalenmittelwerte lieferte somit empirische Belege für die in Kapitel 3 formulierten Ausgangsüberlegungen: Die Wahrnehmung von Lehrkräften in weiterführenden Schulen im Hinblick auf Vergleichsarbeiten erweist sich als deutlich negativer als die von Grundschullehrkräften. Gemäß latenter Mittelwertanalyse empfinden Grundschullehrkräfte im Vergleich zu Lehrkräften weiterführender Schulen VERA demzufolge als nützlicher, haben eine positivere Einstellung und sind eher bereit, mit den erhaltenen Rückmeldungen weiterzuarbeiten.

Auch in den Differenzen der manifesten Skalenmittelwerte spiegeln sich entsprechende Tendenzen wider. Dies trifft auch für das Konstrukt zeitliche Belastung zu, für welches keine latenten Mittelwertdifferenzen betrachtet wurden. Für das Gesamtkonstrukt konstatieren die VERA8-Lehrkräfte insgesamt eine etwas größere empfundene zeitliche Belastung als VERA3-Lehrkräfte. Die manifeste Mittelwertdifferenz erweist sich als auf dem 5 % Niveau signifikant mit einer Effektstärke von $d = 0.10$. Die itemweise Betrachtung der Gruppenunterschiede der zeitlichen Belastung verdeutlichen, dass klare Unterschiede in der Bewertung der einzelnen Aspekte bestehen: Während der Durchführungsaufwand (ZB2) in beiden Gruppen nahezu gleich bewertet wird, empfinden Grundschullehrkräfte den Vorbereitungsaufwand (ZB1) als signifikant ($p < .001$, $d = .0.27$) höher als Lehrkräfte weiterführender Schulen, im Gegensatz zum Auswertungsaufwand (ZB3), der von VERA8-Lehrkräften als deutlich höher beurteilt wird. Bei diesem Item herrscht nicht nur die größte Gruppendifferenz ($d = -0.56$), auch insgesamt wird der Aufwand der Auswertung im Vergleich zu den übrigen Items am höchsten empfunden.

Insgesamt entsprechen die Befunde der latenten Mittelwertanalyse bzw. der ergänzenden Betrachtung manifester Mittelwerte den Erwartungen einer insgesamt im Vergleich positiveren Wahrnehmung von Vergleichsarbeiten in der Grundschule. Lehrkräfte in der Grundschule nehmen eine höhere Nützlichkeit der Vergleichsarbeiten wahr, haben eine positivere Einstellung, konstatieren eine höhere Nutzungsintention und empfinden tendenziell eine geringere zeitliche Belastung im Hinblick auf VERA3 im Vergleich zu Lehrkräften der Sekundarstufe hinsichtlich VERA8.

Untersuchungen zeigen, dass Lehrkräfte in weiterführenden Schulen häufig eine höhere Arbeitsbelastung und Stress empfinden als Lehrkräfte in Grundschulen (vgl. bspw. Antoniou, Ploumpi & Ntalla, 2013; Kavita & Hassan, 2018; Kongcharoen, Onmek, Jandang & Wangyisen, 2020; Roeser et al., 2022). Das erhöhte Stressempfinden zeigt sich auf folgenden Ebenen: Auseinandersetzung mit Eltern und Kolleg*innen, Arbeitsbelastung, beschränkte zeitliche Ressourcen, Eigenschaften der Schüler*innen, Anerkennung und Unterstützung sowie mangelnde Ressourcen (Kavita & Hassan, 2018). Generell unterscheidet sich die Arbeitsrealität zwischen Lehrkräften in Grund- und Sekundarschulen teils deutlich, bspw. durch die Arbeit an größeren Schulen und das Unterrichten verschiedener Klassen an einem Tag (Roeser et al., 2022).

Vor diesem Hintergrund erscheint es durchaus plausibel, dass sich eine höhere Grundbelastung auch in der Bewertung der Vergleichsarbeiten niederschlägt und eben derartige zusätzliche Aufgaben, wie das Schreiben und die Korrektur von Vergleichsarbeiten, von Lehrkräften der Sekundarstufe weniger positiv aufgenommen und bewertet werden als von Grundschullehrkräften. Das höhere Stressempfinden könnte sich somit systematisch in der Beurteilung von Vergleichsarbeiten niederschlagen und entsprechend in einer höheren empfundenen zeitlichen Belastung, geringerer Nutzenwahrnehmung und negativeren Einstellung auswirken und ebenso in einer geringeren Bereitschaft, sich weiter mit den Rückmeldungen auseinanderzusetzen. Insbesondere der Befund einer geringeren Nutzungsbereitschaft deckt sich u. a. mit der Erkenntnis früherer Untersuchungen. Bspw. Demski (2016) ermittelt in Grundschulen eine vergleichsweise höhere Nutzung evidenzbasierter Wissensbestände als in anderen Schultypen.

Neben dieser generellen möglichen Erklärung von Bewertungsunterschieden zwischen Grund- und Sekundarschullehrkräften lohnt sich erneut ein detaillierter Blick auf das Konstrukt zeitliche Belastung. Die einzelnen Items verdeutlichen, dass Lehrkräfte, die VERA8 testen, den Aufwand der Auswertung als deutlich höher empfinden als Lehrkräfte, die mit VERA3 befasst sind. Der Blick in die VERA-Testaufgaben zeigt, dass die Testhefte der VERA8-Testung deutlich umfangreicher sind und somit einen entsprechend höheren Korrekturaufwand erfordern als die VERA3-Tests. Bei VERA8 müssen im Fach Deutsch bspw. je Testheft im Schnitt ca. 100 Items ausgewertet und im Anschluss in eine Eingabemaske übertragen werden, bei VERA3 hingegen nur ca. 50 Items. Entsprechend erklären sich die Unterschiede in der Wahrnehmung der Lehrkräfte. Der etwas höher empfundene Vorbereitungsaufwand (ZB1) könnte

möglicherweise mit der Vorbereitung der Schüler*innen zusammenhängen, da mutmaßlich mehr Zeit beansprucht wird, die jüngeren Kinder in der 3. Klasse auf ein unbekanntes Testformat vorzubereiten als ältere Schüler*innen in der 8. Klasse. Hinsichtlich des Durchführungsaufwandes (ZB2) nimmt die VERA8-Testung in einem Fach zwar tendenziell mehr Zeit in Anspruch, jedoch muss eine Lehrkraft in einer weiterführenden Schule i. d. R. auch nur mit einer Klasse und in einem Fach die Vergleichsarbeiten durchführen, wohingegen eine Grundschullehrkraft häufig VERA3 in Deutsch und in Mathematik mit ihrer Klasse testen muss. Insgesamt nimmt die Testung somit häufig ähnlich viel Zeit in Anspruch, was sich möglicherweise in der vergleichbaren Bewertung des Durchführungsaufwandes niederschlägt.

Fragestellung 3: Kausalanalyse

Im letzten Kapitel des empirischen Teils dieser Arbeit (Kapitel 5.4.4) wurde der Frage nachgegangen, ob sich die Strukturbeziehungen zwischen den Gruppen unterscheiden, also ob Moderatoreffekte der Gruppenzugehörigkeit existieren. Hierfür wurde zunächst ein Gruppenmodell mit zwischen den Gruppen frei geschätzten Pfaden gerechnet und im Anschluss daran ein weiteres Modell, in dem die Pfadkoeffizienten zwischen den Gruppen gleichgesetzt wurden. Beide Modelle wiesen akzeptable (Modell ohne direkten Pfad zwischen Nützlichkeit und Nutzungsintention) bzw. gute (Modell mit direktem Pfad) Kennwerte auf. Wie in den bisherigen Kausalanalysen verbesserte sich die Modellpassung auch im Gruppenmodell deutlich durch die Spezifizierung des direkten Pfades zwischen Nützlichkeit und Nutzungsintention, ein Ergebnis, das erneut die Relevanz der Nutzenwahrnehmung für die Ergebnisnutzung unterstreicht.

Das Modell mit gleichgesetzten Pfadkoeffizienten zeigt im Vergleich zum Modell mit zwischen den Gruppen frei geschätzten Pfaden keine substanzielle Verschlechterung der Modellpassung. Die Ergebnisse sprechen somit dafür, dass sich die Struktur der Pfadkoeffizienten zwischen den Gruppen nicht substanziell unterscheidet und das Kausalmodell für beide Gruppen gültig ist. Der Blick auf die am entsprechenden Modell (Modell 2 $\text{Group}_{(PKg)}$ und Modell 2_{dP} $\text{Group}_{(PKg)}$) geschätzten Modellpfade unterstreicht diesen Befund. Die Kausalstruktur lässt sich in beiden Gruppen weitestgehend abbilden, die Pfadkoeffizenten erweisen sich auf dem 1 % Niveau als signifikant und haben in beiden Gruppen die gleichen Vorzeichen und vergleichbare Effektstärken.

Insgesamt sind die standardisierten Effekte in der VERA8-Gruppe etwas stärker, jedoch i. d. R. lediglich mit einer Differenz von .02 bzw. .03, was sich auch in der etwas höheren Varianzaufklärung der Konstrukte in dieser Gruppe widerspiegelt. Nur der Effekt der zeitlichen Belastung auf die Nutzenwahrnehmung zeigt eine größere Differenz zwischen den Gruppen von .08 ((Modell 2 $Group_{(PKg)}$) und .09 (Modell 2_{dP} $Group_{(PKg)}$)). Dies spricht dafür, dass sich die etwas höhere empfundene zeitliche Belastung durch die Vergleichsarbeiten bei VERA8-Lehrkräften insbesondere in ihrer Nutzenwahrnehmung niederschlägt, teilweise jedoch auch in ihrer Einstellung und Nutzungsintention.

Die Ergebnisse sprechen in der Gesamtbetrachtung für die Generalisierbarkeit der im Modell postulierten Wirkungszusammenhänge. Das untersuchte Kausalmodell erweist sich auch in der Gruppe der Sekundarstufenlehrkräfte im Hinblick auf VERA8 als gültig. Auch die Bedeutung der Nutzenwahrnehmung für die Weiterarbeit mit Vergleichsarbeiten wird in diesem Gruppenmodell erneut deutlich.

6.1.4. Abschließende Betrachtung kritischer Aspekte

Zunächst kann das TAM als Grundlage eines pädagogischen Akzeptanzmodells kritisch hinterfragt werden, da sich die Frage stellt, ob das Modell eine ausreichende Definition von Akzeptanz liefert. Auch das TAM führt den Begriff Akzeptanz lediglich im Titel und nimmt keine konkrete Definition des Akzeptanzbegriffs vor. Die in dieser Arbeit genutzte Definition leitet sich auch vor dem Hintergrund der diskutierten Einstellungstheorien lediglich aus den Modellannahmen und postulierten Konstruktbeziehungen ab. Die empirische Bewährung der Definition wurde daher in Kapitel 5 überprüft. Zudem kann das auf Basis des TAM aufgestellte Forschungsmodell keinen Anspruch auf Vollständigkeit erheben, weil bereits die Aufarbeitung des Forschungsstandes zu Vergleichsarbeiten nahelegt, dass weitere Faktoren existieren, die auf die verschiedenen Facetten des Akzeptanzmodells wirken, in dem aufgestellten Ausgangsmodell jedoch noch keine Berücksichtigung fanden. Daher ist es fraglich, inwiefern das TAM erschöpfend alle Aspekte der Akzeptanz im Kontext von VERA erfasst. Diese Thematik wird in Kapitel 6.4 noch einmal ausführlicher aufgegriffen.

In der empirischen Überprüfung des aufgestellten Modells wurde deutlich, dass ein angepasstes sparsameres Akzeptanzmodell (Modell 2) – ohne die ursprünglich vermutete explizite

Modellierung eines Aufwand-Nutzen-Konstrukts – zu bevorzugen ist und sich in verschiedenen Stichproben bewährt. Es konnte gezeigt werden, dass dieses im Kontext von VERA so wichtige Phänomen der Aufwand-Nutzen-Abwägung in hinreichender Weise durch das Zusammenwirken der Konstrukte zeitliche Belastung und wahrgenommene Nützlichkeit abgebildet werden kann. Eine separate Erfassung und Modellierung scheint zur Erklärung der Einstellung und insbesondere des Nutzungsverhaltens, hier der Nutzungsabsicht, als eines der Hauptziele dieser Arbeit nicht notwendig. Dies spiegelt sich auch in der, im Vergleich der Modelle mit direktem Pfad zwischen Nützlichkeit und Nutzungsintention der Stichprobe 2018 (Modell $1_{dP(V3\text{-}18)}$ und Modell $2_{dP(V3\text{-}18)}$) gleichbleibenden Varianzaufklärung der Nutzungsintention von 64 % wider. Sowohl Modell 1 mit einem separat spezifizierten Aufwand-Nutzen-Konstrukt, als auch Modell 2 ohne diese explizite Konstruktmodellierung, können mit 64 % einen großen Anteil der Varianz der Nutzungsintention der Lehrkräfte aufklären. Anzumerken bleibt, dass die Varianzaufklärung der Einstellung erwartungsgemäß etwas niedriger ist. Durch das Wegfallen des inhaltlich nicht plausiblen extrem hohen Einflusses des Aufwand-Nutzen-Konstrukts von $\beta = .96$ ($p < .001$) in Modell $1_{dP(V3\text{-}18)}$ liegt die Varianzaufklärung in Modell $2_{dP(V3\text{-}18)}$ bei nunmehr 65 % im Vergleich zu 85 % im Ursprungsmodell 1_{dP}. Da der Modellfit von Modell 2 jedoch insgesamt klar bessere Werte aufweist, war dieses zu bevorzugen und wurde für die weiteren Untersuchungen von Gruppenunterschieden zwischen VERA3- und VERA8-Lehrkräften genutzt (siehe Kapitel 5.4). Das dem ursprünglichen TAM ähnlichere Modell passt somit besser zu den Daten und erweist sich in verschiedenen Stichproben als valide.

Auf Basis dieses validierten Modells werden im Folgenden noch einige Details der analysierten Modellkonstrukte beleuchtet. Besonders hervorzuheben ist an dieser Stelle der Einfluss der wahrgenommenen Nützlichkeit. Die Forschungsliteratur zu Vergleichsarbeiten betont vielfach den Zusammenhang von Nützlichkeit und Nutzung bzw. die Bedeutung der Nutzenwahrnehmung als entscheidende Voraussetzung für die Nutzung von und Weiterarbeit mit VERA-Rückmeldungen (siehe bspw. Bonsen et al., 2006; Kühle & Peek, 2007; Vogel, 2020; Wurster, Bach et al., 2016). Auch diese Arbeit unterstreicht die Relevanz der Nutzenwahrnehmung als entscheidender Faktor im Hinblick auf die Nutzung von Vergleichsarbeiten. Die Verbesserung des Modellfits bei der Spezifizierung eines direkten Pfades zwischen Nützlichkeit und Nutzungsintention, sowie die Stärke des entsprechenden direkten Effektes, belegen diese Erkenntnis. Speziell mit Blick auf die Abwägung von Kosten und Nutzen, also der

zeitlichen Belastung und der Nutzenwahrnehmung, wird deutlich, dass der Einfluss der wahrgenommenen Nützlichkeit im Vergleich zu dem der zeitlichen Belastung deutlich überwiegt.

Aus dieser Erkenntnis lässt sich der Schluss ableiten, dass ein zeitlicher (Zusatz-)Aufwand allein kein Hindernis in der Nutzung von Vergleichsarbeiten darstellen sollte, solange Lehrkräfte einen Nutzen für ihren Unterricht ziehen können. Das wichtigste Anliegen zur Förderung von Akzeptanz und Nutzung von Vergleichsarbeiten sollte daher in der in der Steigerung der Nutzenwahrnehmung liegen.

Deshalb lohnt ein kurzer Exkurs zu den modellinhärenten Potenzialen und auch Grenzen potenzieller Maßnahmen, die auf eine Verbesserung der beiden Wahrnehmungsfaktoren zeitliche Belastung und Nutzenwahrnehmung abzielen. Wie bereits ausführlich dargelegt (siehe Kapitel 5.1.3) verdeutlicht die Höhe der Gesamteffekte im finalen Modell $1_{dP(V3\text{-}18)}$ die eindeutige Überlegenheit des Einflusses der Nutzenwahrnehmung gegenüber der zeitlichen Belastung auf die zentrale abhängige Variable Nutzungsintention. Der Gesamteffekt der zeitlichen Belastung auf die Nutzungsintention, also die Summe aller, in diesem Fall indirekten Effekte, liegt bei γ_{ges_NIWN} = -.274 ($p < .001$), der Gesamteffekt der wahrgenommenen Nützlichkeit hingegen bei β_{ges_NIWN} = .775 ($p < .001$).

Auf Basis dieser Gesamteffekte in Kombination mit den manifesten Konstruktmittelwerten und Standardabweichungen lässt sich quantifizieren, inwiefern potenzielle Maßnahmen, die zunächst direkt auf eine Verbesserung der Wahrnehmung der Konstrukte zeitliche Belastung und Nutzenwahrnehmung abzielen, indirekt auch eine Verbesserung der zentralen abhängigen Variablen Nutzungsintention erreichen können. Voraussetzung hierfür ist die Annahme der Gültigkeit des ermittelten finalen Modells. Von Interesse ist hierbei die Frage nach dem Potenzial einer Optimierung des zeitlichen Belastungsempfindens und des wahrgenommenen Nutzens mit Blick auf eine dadurch erzielte Erhöhung der Nutzungsintention sowie die Frage nach den Grenzen einer Einflussnahme. Die Ergebnisse der Berechnung dieser maximal möglichen Einflussnahme auf die Nutzungsintention, jeweils vermittelt über eines der beiden

Konstrukte, sind in Tabelle 50 dargestellt (siehe [10] für die Beschreibung der entsprechenden Berechnungsschritte).

Tabelle 50: Optimierungspotenzial der Konstrukte zeitliche Belastung und Nutzenwahrnehmung

		Zeitliche Belastung	Wahrgenommene Nützlichkeit
a	*M*	3.06	2.49
b	*SD*	0.62	0.74
c	Gesamteffekt auf Nutzungsintention (|*d*|)	0.274	0.775
d	Skalenoptimum	1	4
e	Skalenrange	1-5	1-4
f	Distanz *M* bis $M_{Optimum}$ (Skalenpunkte)	2.06	1.51
g	Distanz *M* bis $M_{Optimum}$ (Standardabweichungen)	3.32	2.04
h	Verbesserung der Nutzungsintention bei $M_{Optimum}$ (Standardabweichungen)	0.91	1.58
i	M_{NI} [a] bei $M_{Optimum}$	3.00	3.46

Anmerkungen. [a] Skala 1-4.

Der Vergleich der für die jeweilige optimale Ausprägung der Konstrukte zeitliche Belastung und Nutzenwahrnehmung ermittelten Werte im Hinblick auf deren Einfluss auf die Nutzungsintention, unterstreicht erneut die deutlich größere Relevanz des Faktors Nutzenwahrnehmung im Vergleich zum zeitlichen Belastungsempfinden. Selbst eine Optimierung des zeitlichen Belastungsempfindens erzielt hinsichtlich ihres Einflusses auf die abhängige Variable Nutzungsintention lediglich eine Verbesserung um 0.91 Standardabweichungen, während eine

[10] Beschreibung des Vorgehens zur Berechnung der maximal möglichen Einflussnahme auf die Nutzungsintention: Ausgehend von den in Kapitel 5.1.2 (siehe u. a. Tabelle 34, S. 169) ermittelten manifesten Konstruktmittelwerten (Zeile a) und Standardabweichungen (Zeile b), wurde zunächst die Distanz in Skalenpunkten (Zeile f) und Standardabweichungen (Zeile g) der jeweiligen Konstruktmittelwerte zum Skalenoptimum (Zeile d) ermittelt. Im nächsten Schritt wurde die Veränderung auf der zentralen abhängigen Variablen, der Nutzungsintention, bei der Annahme der Optimierung der jeweils betrachteten unabhängigen Variablen berechnet. Zum einen wurde diese Veränderungen in der Maßeinheit Standardabweichungen (Zeile h) bestimmt und zum anderen wurde der manifeste Mittelwert der Nutzungsintention bei einer optimalen Ausprägung des manifesten Skalenmittels der jeweilig manipulierten unabhängigen Variablen (Zeile i) errechnet.

Optimierung der wahrgenommenen Nützlichkeit 1.58 Standardabweichungen erreicht. Mit Blick auf das entsprechende theoretische Skalenmittel der Nutzungsintention liegt dieses bei einer optimal geringen zeitlichen Belastung bei $M = 3.00$, bei einer optimal wahrgenommenen Nützlichkeit hingegen bei $M = 3.46$ und liegt somit durch eine Optimierung der Nützlichkeit nur 0.50 Skalenpunkte vom Skalenoptimum von 4 entfernt.

Zentrale Folgerung aus diesen Berechnungen ist, dass Maßnahmen zur Verringerung der zeitlichen Belastung zwar durchaus sinnvoll sind, deren Wirksamkeit jedoch auch klare Grenzen gesetzt sind, wenn es um die Nutzung bzw. Nutzungsintention geht. Eine positive Beeinflussung der Nutzenwahrnehmung bietet größere Potenziale für eine Steigerung der Nutzungsintention. Diese Thematik wird in Kapitel 6.3 zu den Praxisimplikationen noch einmal aufgegriffen.

Ein weiteres Konstrukt, das mit Blick auf die Besonderheiten des finalen Modells noch einmal näher beleuchtet wird, ist die Einstellung, deren Rolle sich als weniger eindeutig erwies als die der Nutzenwahrnehmung: Wirkt die wahrgenommene Nützlichkeit direkt auf die Nutzungsintention, verringert sich der Einfluss der Einstellung auf die Nutzungsintention deutlich bzw. verschwindet je nach Stichprobe. Die Wirkung der Einstellung wird regelrecht überlagert vom Einfluss der wahrgenommenen Nützlichkeit.

Die affektive Beurteilung von Vergleichsarbeiten, repräsentiert durch die Einstellung, scheint somit bei den Lehrkräften bei der Frage nach der geplanten Weiterarbeit mit VERA-Rückmeldungen kaum eine Rolle zu spielen. Die Beurteilung des durch die Vergleichsarbeiten entstehenden Nutzens dominiert dagegen die Entscheidung zur Weiterarbeit mit den rückgemeldeten Daten. Dies könnte auch ein Hinweis darauf sein, dass der Einbezug der Einstellung in das empirische Akzeptanzmodell gar nicht notwendig ist und die Nutzungsintention ausschließlich durch kognitive Prozesse in Form der Beurteilung der Nützlichkeit bzw. die Aufwand-Nutzen-Abwägung bestimmt wird. Die Dominanz des Einflusses der wahrgenommenen Nützlichkeit zur Vorhersage von Verhalten bzw. Handlungsintention gegenüber der Rolle von Einstellung, wird auch in frühen Untersuchungen zum TAM deutlich. In der grundlegenden Arbeit von F. D. Davis et al. (1989) erwies sich die Rolle von Einstellungen zur Erklärung der Beziehung von Überzeugungen und Verhaltensintentionen als nur untergeordnet. Infolgedessen wurde das Konstrukt aus verschiedenen späteren Untersuchungen zum TAM

ausgeschlossen (siehe bspw. Chismar & Wiley-Patton, 2003; Venkatesh & Davis, 2000; Venkatesh, 2000; Venkatesh et al., 2003). Auch diese Modelle erwiesen sich als valide und verweisen auf die Möglichkeit, ggf. auch in zukünftigen Untersuchungen zur Akzeptanz von VERA Modelle ohne eine explizite Einstellungsmodellierung in Betracht zu ziehen, insbesondere mit dem Ziel, auch bei einer Erweiterung des Modells um zusätzliche Konstrukte, ein möglichst sparsames Modell zu erhalten.

Kritisch beleuchtet werden muss außerdem die Operationalisierung des Verhaltensaspekts durch die Intention zur Nutzung der und Weiterarbeit mit den Ergebnisrückmeldungen anstelle einer Erfassung der letztendlich für den Erfolg von VERA relevanten tatsächlichen Ergebnisnutzung. Methodisch ist dies dem Erhebungszeitpunkt, kurz nach der Rückmeldung der Ergebnisse, geschuldet, zu dem noch keine Auseinandersetzung oder Weiterarbeit mit den Rückmeldungen stattgefunden haben kann. Bei einer Datenerhebung zu einem späteren Zeitpunkt, bspw. einige Wochen bis Monate nach der Ergebnisrückmeldung, könnte die zusätzliche Erfassung der tatsächlich stattgefundenen Nutzung weitere interessante Erkenntnisse liefern, u. a. auch über die Konsistenz von Intention und Verhalten. Allerdings existieren Argumente für die in dieser Arbeit gewählte Operationalisierung, da auch Untersuchungen bestätigen, dass die Intention durchaus als adäquater Prädiktor der tatsächlichen Nutzung angesehen werden kann. Turner et al. (2010) ermitteln bspw. auf Basis eines systematischen Literaturreviews Nutzungsintention als besseren Prädiktor der tatsächlichen Nutzung als wahrgenommene Einfachheit und Nutzenwahrnehmung. Diese Erkenntnis spricht u. a. dafür, Nutzungsintention in jedem Fall im Akzeptanzmodell zu integrieren und wenn möglich die tatsächliche Nutzung zusätzlich zu berücksichtigen, um weitere Erkenntnisse über den Nutzungsprozess und, diesem nachgelagert, eine angestrebte Unterrichtsentwicklung zu erhalten.

Ein weiterer Punkt, der an dieser Stelle noch einmal bedacht werden soll, betrifft die Operationalisierung der zeitlichen Belastung anstelle von Einfachheit im untersuchten Akzeptanzmodell. Die Entscheidung, die wahrgenommene Einfachheit durch die Bewertung des Zeitaufwandes zu subsituieren, war zum einen inhaltlichen und zum anderen methodischen Überlegungen geschuldet. Unter inhaltlichen Gesichtspunkten ist die Einfachheit schwierig eindeutig zu operationalisieren. Zunächst müsste eine Definition für das Konstrukt Einfachheit im Kontext von Vergleichsarbeiten gefunden werden. Hierfür gäbe es mehrere Optionen: Zum einen könnte sich Einfachheit auf die Einfachheit bzw. Verständlichkeit der einzelnen Stufen des Durchführungsprozesses beziehen, und zum anderen auch auf die Einfachheit der

Darstellungen bzw. die Verständlichkeit der Rückmeldungen. Insgesamt würde eine entsprechend umfangreiche Konstruktoperationalisierung die Komplexität des Modells deutlich erhöhen und das Phänomen der zeitlichen Belastung wäre noch nicht abgebildet. Hinzu kommt aus methodischer Sicht, dass zum Erhebungszeitpunkt, direkt nach der Dateneingabe, die Einfachheit der Ergebnisrückmeldungen noch nicht beurteilt werden konnte, weshalb dieser Aspekt der Einfachheit nicht hätte erfasst werden können. Aufgrund dieser datenerhebungsbegründeten Einschränkungen, und um das Modell möglichst sparsam zu halten, wurde die Substituierung des Konstrukts Einfachheit durch die zeitliche Belastung gewählt.

Im Hinblick auf das Konstrukt zeitliche Belastung bleibt an dieser Stelle weiterhin zu hinterfragen, ob das Konstrukt in der genutzten Operationalisierung bereits vollumfänglich alle relevanten Aspekte berücksichtigt. Denkbar wäre bspw. eine Erfassung des Zeitaufwandes zur Dateneingabe oder der Auseinandersetzung mit den Ergebnissen, im Sinne der bspw. in den Arbeiten von Koch (2011) und Groß Ophoff (2013) operationalisierten Rezeption als separates Konstrukt. Bisher bleibt es der Interpretation der Lehrkräfte überlassen, ob sie den Aufwand der Dateneingabe in die Bewertung des Auswertungsaufwandes miteinbeziehen. Es bleibt zu vermuten, dass einige dies tun, was zu einer Verzerrung des Items führen könnte. Zukünftige Forschungsarbeiten sollten diese Aspekte bei der Konstruktoperationalisierung in Betracht ziehen.

6.2. Limitationen

Bei der Bearbeitung der dieser Arbeit zugrundeliegenden Fragestellung wurden verschiedene Einschränkungen deutlich, die bspw. auf die Auswahl der Theorie, die gewählte Datengrundlage oder Methode zurückzuführen waren. Die meisten dieser Limitationen wurden im Verlauf der vorliegenden Arbeit bereits ausführlich behandelt und der Umgang mit den aufgetretenen Problemen detailliert dokumentiert, weshalb sich dieses Kapitel im Wesentlichen, auf die bisher noch nicht behandelten Problematiken beschränkt, die in den folgenden Abschnitten beschrieben werden.

Zu beachten sind dabei zunächst ganz grundsätzliche Limitationen von Survey-Studien: Die erhobenen Daten und somit die vorgestellten Ergebnisse dieser Arbeit beruhen auf den Selbstauskünften der befragten Lehrkräfte, einer Erhebungsmethode, bei der generelle Verzerrungen

in den gewonnenen Daten vorliegen können und deren Validität zumindest hinterfragt werden kann. Generelle Einschränkungen bei der Erhebung von Selbstauskünften beziehen sich dabei bspw. auf die Tendenz zu sozialer Erwünschtheit oder individuell unterschiedliche Referenzrahmen und Interpretationen bei der Beantwortung der Fragen. Durch die Verwendung verbaler Ratingskalen (Likert-Skala) wurde jedoch durch den dadurch gebotenen interpersonell vergleichbaren Maßstab eine vergleichbare Interpretation der Antwortoptionen durch die verschiedenen Proband*innen gewährleistet werden (siehe bspw. Jonkisz, Moosbrugger & Brandt, 2012). Auch Vogel (2020) spricht diese Problematik in seiner Arbeit an, verweist jedoch darauf, dass trotz aller Einschränkungen die Verwendung von Likert-Skalen in der empirischen Bildungsforschung ein etabliertes Erhebungsinstrument darstellt, zu dem kaum vergleichbare Alternativen bestehen, gerade im Hinblick auf eine Einstellungsmessung bei Lehrkräften.

Dem Phänomen des sozial erwünschten Antwortverhaltens sollte mit dem Hinweis auf Anonymität bei der Datenerhebung bzw. anonyme Datenhaltung entgegengewirkt werden. Das teils sehr negative Antwortverhalten der Lehrkräfte lässt vermuten, dass soziale Erwünschtheit im Falle dieser Arbeit kein großes Problem dargestellt hat. Der Gefahr unterschiedlicher Interpretationen von Items wurde versucht, mit möglichst klaren Formulierungen zu begegnen. Wie bereits angesprochen, hätte bspw. Item ZB3 („Wie bewerten Sie den Zeitaufwand für die Auswertung?“) jedoch eindeutiger formuliert werden können. Hier bleibt Raum für Interpretation, ob dieses Item auf die reine Korrekturzeit oder auch auf die Übertragung der Daten in das VERA-Portal bezogen ist. Die Formulierung für zukünftige Erhebungen sollte nachgebessert werden.

Eine weitere Einschränkung dieser Arbeit betrifft die Repräsentativität der gewonnenen Erkenntnisse aufgrund der Beschaffenheit der Stichprobe. Die Freiwilligkeit der Teilnahme und fehlende Randomisierung der Stichprobenauswahl schränken die Generalisierbarkeit der Ergebnisse ein. Dies könnte zu Verzerrungen, bspw. durch die verstärkte Teilnahme besonders motivierter oder auf der anderen Seite sehr verärgerter Lehrkräfte, geführt haben. Insgesamt spricht jedoch die Reproduzierbarkeit der Ergebnisse mit verschiedenen Stichproben, sowohl auf Mess- als auch auf Strukturebene, eher für die Belastbarkeit der Befunde. Generell könnte auch die Rücklaufquote limitierend für die Generalisierbarkeit der Ergebnisse wirken. Jedoch verrät der Blick auf ähnliche Untersuchungen, wie bereits in Kapitel 4.2.2 dargelegt, dass Rücklaufquoten zwischen 18 % und 39 %, wie sie in dieser Arbeit vorliegen, durchaus

erwartbar sind und i. d. R. bei Studien zu Vergleichsarbeiten Rücklaufquoten von rund 25 % durchaus üblich sind (Wacker & Kramer, 2012).

Des Weiteren bleibt anzumerken, dass ähnlich wie bspw. bei Wagner und Koch (2021) aus datenschutzrechtlichen Gründen keine persönlichen Angaben wie Alter oder Geschlecht sowie keine sonstigen Kontextinformationen erhoben werden konnten, weshalb dahingehend keine weiteren Detailanalysen durchgeführt werden konnten. Jedoch finden sich auch in der verwandten Literatur keine Hinweise auf derartige Einflüsse, weshalb diesbezüglich auch keine expliziten starken Hypothesen bestanden. Während Kontextvariablen in den meisten Untersuchungen mit Lehrkräften zu Vergleichsarbeiten unberücksichtigt bleiben, zeigen u. a. Ditton et al. (2002), dass bspw. das Geschlecht keine Rolle bei Einstellungen von Lehrkräften zu zentralen Tests spielt.

Zudem ist der Erhebungszeitpunkt direkt nach der Ergebnisrückmeldung ein neuralgischer Punkt dieser Arbeit. Optimalerweise sollten die Lehrkräfte zu einem späteren Zeitpunkt (erneut) befragt werden, nachdem bereits eine Auseinandersetzung mit den Ergebnissen der Vergleichsarbeiten stattgefunden hat: Zum einen, um dann besser die Nützlichkeit bewerten zu können, zum anderen, um auch eine Bewertung der Verständlichkeit der Rückmeldungen erheben und neben der Nutzungsintention die tatsächliche Nutzung untersuchen zu können. In anderen Untersuchungen mit ähnlichen Voraussetzungen zeigt sich jedoch, dass ein früher Erhebungszeitpunkt keine allzu große Einschränkung darstellt. Maier (2008b) bspw. beschreibt ein ähnliches Problem: In seiner Untersuchung fand die Datenerhebung in Form einer schriftlichen Befragung auch kurz nach bzw. mit Erhalt der Rückmeldungen statt, was zu Messfehlern und Verzerrungen führen kann, insbesondere im Hinblick auf Fragen zur schulinternen Auseinandersetzung und Diskussion über die Ergebnisse, die u. U. erst deutlich später stattfindet. Der Autor gelangt jedoch zu dem Schluss, dass Lehrkräfte die Situation an ihrer Schule und Diskussionen über Testergebnisse auch bei einem frühen Bearbeitungszeitpunkt des Fragebogens gut antizipieren können und häufig auch auf Erfahrungswerte der vergangenen Schuljahre zurückgreifen können, sodass der Erhebungszeitpunkt, wenn überhaupt, nicht zu übermäßigen Verzerrungen führen sollte.

Auch im Hinblick auf die dieser Arbeit zugrundeliegenden Datenerhebung kann von der Annahme ausgegangen werden, dass die befragten Lehrkräfte sowohl Nützlichkeit als auch die

anderen Wahrnehmungsaspekte entweder aus ihren eigenen Erfahrungen der vorangegangenen Durchgänge ableiten oder basierend auf kollektiven Erfahrungen und dem Austausch im Kollegium gut antizipieren können und somit befähigt waren, den Fragebogen adäquat zu beantworten.

Abschließend bleibt zu betonen, dass es sich bei den in dieser Arbeit untersuchten Daten um eine Querschnittserhebung handelt, weshalb keine Aussagen zu tatsächlichen kausalen Wirkungszusammenhängen gemacht werden können, sondern lediglich korrelativ signifikante Zusammenhänge untersucht werden. So konnte nicht eindeutig geklärt werden, ob eine positive Einstellung gegenüber Vergleichsarbeiten tatsächlich kausal durch einen großen wahrgenommenen Nutzen bedingt wird.

Mit der Frage nach den Implikationen der bisher diskutierten theoretischen und empirischen Befunde dieser Arbeit für den operationalen Umgang mit Vergleichsarbeiten in der Praxis, beschäftigt sich das folgende Kapitel 6.3.

6.3. Praxisimplikationen

Die Ergebnisse dieser Arbeit geben Aufschluss über die Wahrnehmung von Vergleichsarbeiten durch Lehrkräfte und verdeutlichen dadurch, an welchen Stellen noch Verbesserungspotenziale bestehen, und wo durch die verantwortlichen Institutionen nachgebessert werden sollte. Insbesondere die Relevanz der wahrgenommenen Nützlichkeit legt nahe, dass es für die erfolgreiche Nutzung von Vergleichsarbeiten eines der wichtigsten Ziele sein muss, eine positive Nutzenerfahrung für Lehrkräfte zu schaffen und die Nutzenwahrnehmung zu verbessern. Darüber hinaus stellen auch die weiteren untersuchten Aspekte mögliche Ansatzpunkte zur Steigerung der Akzeptanz und zur Förderung einer Weiterarbeit mit VERA-Rückmeldungen dar. Hier verdeutlichen die manifesten Konstruktmittelwerte mit ihren insgesamt negativen Tendenzen (siehe auch Tabelle 34, S. 180; Tabelle 40, S. 198; Tabelle 47, S. 218), dass hinsichtlich aller untersuchten Facetten ein großes Verbesserungspotenzial besteht. Hier wären verschiedene Stellschrauben denkbar, die in diesem Kapitel exemplarisch erläutert werden.

Zunächst lohnt sich ein Blick auf die in den letzten Jahren initiierten zentralen Neuerungen bei den Vergleichsarbeiten: die schrittweise Implementierung einer computerbasierten Testung (CBT) und die Modularisierung der Testaufgaben, und deren mögliche Auswirkungen auf die Wahrnehmung der Lehrkräfte. Hierbei stellt der Übergang von der bisher noch dominierenden papierbasierten Testung zu einer computerbasierten Testung den vermutlich wirkungsvollsten Ansatzpunkt zur Verringerung der zeitlichen Belastung dar. Durch eine computerbasierte Testung entfällt der, zumindest in einigen Bundesländern, zum Zeitpunkt der Datenerhebung noch notwendige Schritt der Vorbereitung der Testmaterialien (z. B. Druck der Testhefte). In jedem Fall verringert sich jedoch der Auswertungsaufwand, der häufig in Befragungen bemängelt wird, deutlich. Die Korrektur geschlossener Items entfällt bei einer CBT vollständig, sodass nur offene Items weiterhin von der Lehrkraft ausgewertet müssen. Die aufwendige Übertragung der Ergebnisse in das VERA-Portal erübrigt sich ebenfalls. Die Vermutung wäre, dass durch die erzielte Zeitersparnis und das dadurch verringerte Belastungsempfinden die Nutzung der Ergebnisse indirekt über die im Modell implizierten Pfade gefördert werden kann. Dies könnte insbesondere der Fall sein, wenn die Durchführung der Vergleichsarbeiten im bereits ausgelasteten Schulalltag nicht mehr als so große Belastung wahrgenommen wird.

Erste Hinweise auf ein verringertes Belastungsempfinden bei einer computerbasierten Testung im Vergleich zur Paper-Pencil (PP) Testung, besonders im Hinblick auf einen deutlich niedrigeren Auswertungsaufwand, zeigen sich in nicht veröffentlichten Lehrkräftebefragungen des zepf im Zuge der schrittweisen Einführung computerbasierter VERA-Testungen. Auch in Tabelle 13 (siehe S. 128) sind diese Tendenzen erkennbar, die signifikanten Mittelwertunterschiede und die ausgeprägte Effektstärke bei Item ZB3 (Zeitaufwand der Auswertung) im Vergleich der betrachteten CBT und PP Lehrkräfte verdeutlichen, dass die faktische Zeitersparnis bei der Korrektur, durch eine computerbasierte Testung, sich auch in der Wahrnehmung der Lehrkräfte niederschlägt. Belastbare empirische Belege hierzu und zu den weiteren potenziellen Auswirkungen eines veränderten Testmodus auf die weiteren Modellfacetten müssen jedoch zukünftige Untersuchungen erbringen (siehe Kapitel 6.4).

Mit dem Ziel, die Nutzung von Vergleichsarbeiten zu fördern, sollte es, gemäß den Erkenntnissen der vorliegenden Arbeit, das zentrale Anliegen der Verantwortlichen sein, die Nutzenwahrnehmung von Lehrkräften zu erhöhen, da sich diese als wichtigster Einflussfaktor der

Nutzung der rückgemeldeten Daten herausgestellt hat. Hierbei sind u. a. die beiden folgenden Ansatzpunkte denkbar: zum einen können die Testmodalitäten in den Blick genommen werden und zum anderen die rückgemeldeten Daten. Hinsichtlich der Testmodalitäten besteht mit der Modularisierung, die im Jahr 2012 im Zuge der Vereinbarung zur Weiterentwicklung von VERA von der KMK verabschiedet wurde (siehe KMK, 2018), bereits ein Ansatz, die Testhefte stärker an die individuellen Bedürfnisse der Lehrkräfte bzw. der jeweiligen Schüler*innen anzupassen. Der Gedanke hinter dieser Modularisierung ist eine flexiblere bedarfsorientierte Auswahl an Testheften, sprich eine individuelle Festlegung von Kompetenzbereichen und -niveaus. Die Festlegung der Entscheidungsebene erfolgt je Bundesland, möglich ist eine Entscheidung auf Landes-, Schul-, Klassen- und sogar Schüler*innenebene. Konkret wird jeweils ein Basismodul verpflichtend getestet, ein Ergänzungsmodul kann entsprechend der Landesvorgaben auf der jeweils definierten Ebene frei gewählt werden.

Die Modularisierung setzt an einem Kritikpunkt am Inhalt der Vergleichsarbeiten an, der einen möglichen negativen Einflussfaktor der Nutzenwahrnehmung darstellen könnte, der mangelnden Passung der Testinhalte zu den bisher behandelten Unterrichtsthemen. Lehrkräfte kritisieren hierbei eine mangelnde curriculare Passung der Vergleichsarbeiten, da sich deren Inhalte an Bildungsstandards und nicht an den Lehrplänen der Schulen orientieren (siehe bspw. Diemer, 2013; Kuper et al., 2016; F. Thiel et al., 2019). Dies führt dazu, dass Inhaltsbereiche getestet werden, die im Unterricht noch nicht thematisiert wurden und somit Aufgaben gestellt werden, die die Schüler*innen noch nicht beantworten können, was sich nach Aussagen von Lehrkräften bspw. in VERA begleitenden Evaluationsbefragungen negativ auf die Motivation der Schüler*innen auswirken kann. Durch eine flexible Zusammenstellung von Testheften bis hin zu einer binnendifferenzierten Auswahl von Testaufgaben könnte diesem Problem entgegenwirkt werden, da die Testhefte stärker an die individuellen Voraussetzungen in den Klassen bzw. der Schüler*innen angepasst werden könnten. Davon wäre eine positive Wirkung auf die Nutzenwahrnehmung zu erhoffen, da aus Sicht der Lehrkräfte die Rückmeldungen ggf. aussagekräftiger wären. Vermutlich erachten Lehrkräfte dagegen Rückmeldungen zu Testheften mit Aufgaben, die die meisten Schüler*innen noch gar nicht lösen können, da Themen noch nicht behandelt wurden, als eher nutzlos und sind entsprechend nicht bereit, diesen weitere Beachtung zu schenken. In diesem Zusammenhang sollten unbedingt auch Möglichkeiten adaptiver Testung bzw. *Multistage-Testing* in Betracht gezogen werden. Hierbei erfolgt die Auswahl von einzelnen Testaufgaben bzw. Testteilen basierend auf der korrekten Beantwortung der vorangegangenen Aufgaben und ermöglicht eine passgenaue, auf die individuellen

Leistungsniveaus der Schüler*innen angepasste Testung (Berger & Moser, 2020). Diese Art der Testung gestaltet sich, ebenso wie die generelle Zusammenstellung der Testhefte, bei einer computerbasierten Testung deutlich einfacher als bei der klassischen Paper-Pencil-Testung.

Zudem sollte der gesamte Prozess der Vergleichsarbeiten nicht nur inhaltlich, sondern auch zeitlich stärker flexibilisiert werden, um den individuellen Anforderungen häufig ganz unterschiedlicher Schulrealitäten gerecht zu werden. Denkbar wären in diesem Zusammenhang auch, auf individuelle Bedürfnisse von Schulen, Klassen und Schüler*innen abgestimmte, Re-Tests, bspw. um Feedback über Erfolge von Unterrichtsentwicklungsmaßnahmen zu erhalten oder die Leistungsentwicklung infolge eines im VERA-Test aufgedeckten Defizits einer Klasse oder einzelner Schüler*innen zu monitoren.

Bislang lässt die Umsetzung von Flexibilisierung und Modularisierung jedoch noch viel Raum für Optimierung. Zeitliche Flexibilisierung gibt es in den meisten Bundesländern bisher nur insofern, als dass weiterhin zeitlich begrenzte Testzeitfenster angeboten werden, innerhalb derer die Vergleichsarbeiten in den Schulen durchgeführt werden können. Denkbar wäre jedoch durchaus auch die Möglichkeit eine Testung über das gesamte Schuljahr, je nach zeitlichen Kapazitäten und Testbedarfen in den Schulen.

Auch im Hinblick auf bisherige Gestaltung einer modularisierten Testung lässt sich resümieren, dass diese noch nicht allzu umfassend realisiert wurde. Die Bedürfnisse von Schulen werden hierbei kaum in den Blick genommen und es gibt nur minimale Wahlmöglichkeiten, die, wenn vorhanden, häufig bereits auf Schulebene, deutlich seltener auf Klassen- oder gar auf individueller Ebene, getroffen werden. Schulen bzw. Lehrkräfte haben ggf. andere bzw. spezifischere Fragen zum Leistungsstand ihrer Schüler*innen, die zwar grundsätzlich mit einem externen Evaluationsinstrument wie VERA gut zu beantworten wären, jedoch u. U. nicht in der aktuellen Umsetzung. Bspw. interessiert sich eine Lehrkraft aus individuellen Gründen speziell für den Leistungsstand ihrer Klasse in der Domäne Zahl, jedoch weniger für die Domäne Raum und Form, welche jedoch im betreffenden Durchgang getestet wird. In diesem Falle liefert die VERA-Testung keine Antworten auf die eigentlichen Fragen der Lehrkraft, was vermutlich zu Ablehnung und einer Nicht-Nutzung der Ergebnisse führen wird. Hätten die Vergleichsarbeiten hingegen Informationen zum Leistungstand der eigentlich interessierenden Domäne geliefert, wären diese vermutlich auf deutlich positivere Resonanz gestoßen

und hätten gemäß dem untersuchten Modell zu einer höheren Nutzenwahrnehmung und auch einer Nutzungsintention oder eben tatsächlichen Nutzung der Ergebnisse geführt.

Neben einer Anpassung der Testmodalitäten durch Modularisierung und Flexibilisierung, sollten die zur Weiterarbeit vorgesehenen rückgemeldeten Daten in den Blick genommen werden. In verschiedenen Forschungsarbeiten wird in diesem Zusammenhang die Wichtigkeit einer Anschlussfähigkeit der Ergebnisrückmeldungen an den Unterricht bzw. die schulische Realität diskutiert (siehe bspw. Altrichter et al., 2016; Bez, Poindl, Bohl & Merk, 2021; Wurster, Bach et al., 2016). Diese ist häufig geprägt von begrenzten zeitlichen und personellen Ressourcen. Hier wäre es notwendig, dass sich die Rückmeldungen bzw. die Weiterarbeit mit diesen ohne großen zeitlichen Aufwand in die Unterrichtsprozesse integrieren lassen. Dies schont zum einen zeitliche Ressourcen und könnte zum anderen möglicherweise die Nutzenwahrnehmung der Lehrkräfte erhöhen, wenn sie in den Rückmeldungen unterrichtsrelevante, verständliche Informationen erkennen, die ihnen sinnvolle ergänzende Hinweise und ggf. auch konkrete Materialien für die weitere Unterrichtsgestaltung liefern. Hierbei wünschen sich Lehrkräfte häufig eher Informationen auf Individualebene, in übersichtlicher (grafischer) Darstellung, die auch mit rudimentären statistischen Kenntnissen leicht verständlich sind, und darüberhinausgehende Informationen zur Interpretation, konkreten Nutzung und Weiterarbeit im Unterricht (siehe bspw. Altrichter et al., 2016; Schneewind & Kuper, 2009; Skejic et al., 2015).

Zusammenfassend kann die Empfehlung ausgesprochen werden, zunächst die bereits beschlossenen Maßnahmen zur Weiterentwicklung der Vergleichsarbeiten voranzutreiben und deren Potenziale voll auszuschöpfen. Besonders der Ausbau der computerbasierten Testung erleichtert dahingehend die Umsetzung vieler Neuerungen und kann somit förderlich auf verschiedene kritische Aspekte in der Beurteilung von Vergleichsarbeiten bei Lehrkräften einwirken. Sowohl Modularisierung, einschließlich möglicher adaptiver bzw. *Multistage-Testing*, als auch innovative Rückmeldeformate, möglicherweise inklusive bspw. interaktiver Materialien zur Weiterarbeit, lassen sich digitalisiert deutlich leichter umsetzen und können Lehrkräften neue Optionen zur Arbeit mit VERA anbieten. Hinzu kommt der direkte Einfluss der Zeitersparnis vor allem bei der Auswertung, der sich in der Bewertung der Lehrkräfte positiv auswirken sollte.

Bei der Umsetzung entsprechender Maßnahmen sollte jedoch die zentrale Erkenntnis dieser Arbeit berücksichtigt werden, dass Nutzenwahrnehmung den weit bedeutenderen Einflussfaktor darstellt und der Einfluss einer zeitlichen Entlastung für Lehrkräfte limitiert und ab einem gewissen Punkt gänzlich erschöpft ist. Wie durch die ergänzenden Erkenntnisse in Kapitel 6.1.4 (siehe bspw. Tabelle 50) herausgestellt, ist bspw. die positive Wirkung eines veränderten Testmodus bspw. durch die Umstellung auf eine computerbasierte Testung vermutlich limitiert, da diese vor allem auf die zeitliche Belastung und kaum auf die Nutzenwahrnehmung wirkt. Während sich eine Reduzierung des wahrgenommenen Aufwands in ihrer Wirkung als begrenzt erweist, stellt sich die Nutzenwahrnehmung als klar effektivere Stellschraube heraus. Von einer zu starken Fokussierung auf die Stellschraube der zeitlichen Entlastung sollte daher abgesehen werden, zumal der Beschluss zur Weiterentwicklung der Vergleichsarbeiten generell eine umfassende Umstellung von einer papierbasierten hin zu einer computerbasierten Testung vorsieht (KMK, 2018). Daher ist anzunehmen, dass die Potenziale möglicher zeitlicher Entlastungen in den kommenden Jahren ohnehin vollständig ausgeschöpft werden.

Aus diesem Grund sollte der Fokus weiterer Maßnahmen mit dem Ziel einer Akzeptanz- und Nutzungsförderung von Vergleichsarbeiten auf Seiten der schulischen Akteur*innen, verstärkt auf einer Verbesserung der Nutzenwahrnehmung liegen. Bspw. liefern u. a. erste Ergebnisse einer bisher unveröffentlichten Interviewstudie zur Nutzung von Vergleichsarbeiten in Schulen (WeSU) Hinweise darauf, was schulische Akteur*innen im Hinblick auf Vergleichsarbeiten als nützlich erachten bzw. was sie sich mit Blick auf Vergleichsarbeiten wünschen würden. Hierbei wird u. a. der Wunsch nach intensiverer Unterstützung sowie einer Einbettung in ein übergreifendes System deutlich. Möglicherweise könnte dies sinnvolle Ansatzpunkte darstellen, um die zentrale Stellschraube Nutzenwahrnehmung weiterzudrehen.

6.4. Forschungsimplikationen und -ausblick

Auf der theoretischen Ebene leistet die vorliegende Arbeit einen Beitrag zur Akzeptanzforschung im Kontext der empirischen Bildungsforschung, konkret im Hinblick auf das Verständnis der Akzeptanz von Vergleichsarbeiten aus der Sicht von Lehrkräften. Wurde bisher in der Literatur mit einem undifferenzierten Akzeptanzbegriff gearbeitet, handelt es sich bei dieser Arbeit um die erste, die sich der Frage der Akzeptanz im Kontext von VERA aus einer

theoretischen Perspektive annähert und eine klare Begriffsdefinition aufstellt. Zeichnete sich das Forschungsfeld bisher durch einen fehlenden Konsens und fehlende Struktur hinsichtlich Definition und Konzeptualisierung von Akzeptanz und der damit einhergehenden Behinderung von Forschungsfortschritt und Erkenntnisgewinn aus, ermöglichen die Definition von Akzeptanz sowie die theoriegeleitete Modellkonzeptualisierung weiteren Untersuchungen an die Erkenntnisse dieser Arbeit anzuschließen und die Akzeptanzforschung zu Vergleichsarbeiten voranzubringen.

Zentraler Erkenntnisgewinn liegt hierbei in der Herausarbeitung der Bedeutung von Nützlichkeit, insbesondere gegenüber der, in der Literatur häufig zitierten, zeitlichen Belastung. Das in dieser Arbeit validierte Modell ermöglicht eine Quantifizierbarkeit des Einflusses dieser beiden Faktoren und zeigt zudem die Grenzen dieser beiden potenziellen Stellschrauben im Sinne einer externen Einflussnahme auf Akzeptanz und Nutzung von Vergleichsarbeiten auf. Zu bemerken ist hierbei, dass sich eine separate Konstruktmodellierung der in der Forschungsliteratur viel betonten Abwägung von Aufwand und Nutzen als nicht notwendig erwies, sondern sich über das Zusammenspiel der einzelnen Konstrukte Nützlichkeit und zeitliche Belastung hinreichend abbilden lässt.

Des Weiteren leistet diese Arbeit einen Beitrag zur Ausweitung der TAM-Forschung, indem dieses bisher nur im Bereich von Informations- bzw. Kommunikationstechnologien eingesetzte Modell in adaptierter Fassung in das Gebiet der empirischen Bildungsforschung übertragen wurde. Die Überprüfung des aufgestellten Modells zeigt, dass das TAM, in angepasster Version auch im nicht-technischen Bereich anwendbar ist und dass auch im untersuchten Kontext der Vergleichsarbeiten, ebenso wie im ursprünglichen Anwendungsbereich, die Nutzenwahrnehmung den wichtigsten Einflussfaktor der Nutzung darstellt.

Durch die Definition von Akzeptanz und die daraus abgeleitete Modellierung des Akzeptanzprozesses auf Basis des TAM leistet diese Arbeit einen Beitrag zur Erklärung der Nutzung bzw. Nicht-Nutzung von Vergleichsarbeiten. Jedoch kann sicherlich kein Anspruch auf Vollständigkeit des untersuchten Modells erhoben werden. Sowohl die bisherige TAM-Forschung als auch Forschungsarbeiten sowie Praxiserfahrungen zu Vergleichsarbeiten legen nahe, dass verschiedene weitere Aspekte existieren, die in zukünftigen Untersuchungen Berücksichtigung finden sollten. Abschließend soll daher unter Berücksichtigung verbleibender

Forschungsdesiderate im Folgenden ein Konzeptentwurf für zukünftige Forschungsarbeiten skizziert werden.

Diese Konzeption, die in Abbildung 25 schematisch dargestellt ist, verfolgt drei Ansatzpunkte: (1) Erweiterung des untersuchten Akzeptanzmodells um weitere Wahrnehmungsfacetten bzw. eine Detailanalyse des als zentral identifizierten Faktors der Nutzenwahrnehmung; (2) Untersuchung des Einflusses verschiedener externer Faktoren (Inputfaktoren); (3) Wirkung von Akzeptanz im Sinne einer Untersuchung des Outputs bzw. Outcomes von Vergleichsarbeiten. Während Punkt (1) die Ausweitung des untersuchten Modells bzw. die weitere Ausdifferenzierung der Konzeption und Definition von Akzeptanz betrifft, erfolgt mit den Punkten (2) und (3) eine Einbettung des Modells in einen größeren Zusammenhang, indem nicht nur Inputfaktoren und äußere Gegebenheiten, die die Wahrnehmung von Lehrkräften potenziell beeinflussen können, berücksichtigt werden, sondern auch die Folgen einer Weiterarbeit mit Vergleichsarbeiten.

Zwar wurde in der vorliegenden Arbeit eine theorie- und empiriebasierte Definition bzw. Konzeption von Akzeptanz im Kontext von VERA erarbeitet und validiert, dennoch stellt sich die Frage, ob diese Konzeption bereits vollumfänglich alle Aspekte von Akzeptanz umfasst oder ob noch weitere Faktoren existieren, die die Beurteilung von Vergleichsarbeiten beeinflussen. In Abbildung 25 ist dieser Teil (1) möglicher Modellerweiterungen in dem äußeren gestrichelten Kasten dargestellt, welcher die Akzeptanz, gemäß der in dieser Arbeit verwendeten Konzeption und die entsprechend stattfindenden Wahrnehmungs- und Verarbeitungsprozesse repräsentiert.

Nachdem im Rahmen der vorliegenden Arbeit ein Akzeptanzmodell gefunden und validiert wurde, wäre es im nächsten Schritt sinnvoll, die im Modell untersuchten einzelnen Einflussfaktoren, besonders mit Blick auf die Nutzenwahrnehmung, im Detail weiter zu beleuchten, um herauszufinden, was konkret getan werden kann, um diese Stellschrauben zu drehen. Nachdem die Nutzenwahrnehmung als wichtigster Einflussfaktor identifiziert wurde, wäre es nun forschungslogisch notwendig, aufzuschlüsseln wie diese auf Seiten der Lehrkräfte zustande kommt. Hierfür scheint eine Survey-Studie jedoch wenig geeignet, sodass es aus forschungstheoretischer Sicht ratsam wäre, einen methodischen Paradigmenwechsel anzustreben und, bspw. mit Hilfe von Interviews mit schulischen Akteur*innen, diese Details und Facetten

näher zu beleuchten. Hierbei steht die forschungsleitende Frage im Vordergrund, was Nützlichkeit für Lehrkräfte in diesem Kontext im Detail bedeutet und wie diese verbessert werden kann. Erkenntnisse einer solchen qualitativen Untersuchung lassen sich u. U. wiederum für eine weiterführende modellerweiternde quantitative Studie nutzen, indem sich Hinweise auf Operationalisierungen neuer Konstrukte oder für angepasste Konstruktoperationalisierungen ergeben. Die bereits zum Ende des Kapitels 6.3 erwähnte Interviewstudie WeSU stellt bspw. einen ersten Ansatz einer solchen Untersuchung dar.

Mit Blick auf erwägenswerte Modellerweiterungen sollte zudem u. a. der Nutzungsaspekt näher unter die Lupe genommen werden. In der vorliegenden Arbeit konnte aufgrund des Erhebungszeitpunktes nur die Intention zur Nutzung erhoben werden, nicht die tatsächlich stattgefundene Weiterarbeit mit den rückgemeldeten Daten. Diesen Aspekt sollten künftige Forschungsarbeiten dringend berücksichtigen. Zwar gilt eine konstatierte Nutzungsintention gemäß dem TAM als valider Prädiktor des tatsächlichen Verhaltens (siehe bspw. Turner et al., 2010), jedoch wurde das TAM bisher noch nicht im Bereich der empirischen Bildungsforschung generell oder speziell im Kontext von Vergleichsarbeiten genutzt. Daher sollte die Eignung der Nutzungsintention zur Prognose der tatsächlichen Nutzung und somit zur Operationalisierung des Nutzungsaspektes empirisch überprüft werden.

Darüber hinaus liefern sowohl die TAM-Forschung als auch die Forschungsliteratur zu Vergleichsarbeiten Hinweise auf mögliche weitere relevante Gesichtspunkte. Hier liefert Kapitel 2.2.2.6 bereits, basierend auf Forschungsergebnissen zur Wahrnehmung von VERA, verschiedene Ansatzpunkte zu möglichen Modellerweiterungen. Auch in der Forschungsliteratur zum TAM existieren bereits viele Arbeiten, die sich mit der Untersuchung von Modellerweiterungen beschäftigen, von denen sich einige auch im Kontext von VERA wiederfinden lassen. In der ursprünglichen Version des TAM werden Konstrukte wie subjektive Normen, die bspw. in der TRA einen weiteren Prädiktor der Verhaltensintention darstellen, zunächst nicht berücksichtigt. Arbeiten zur Weiterentwicklung des Modells wie bspw. die von Venkatesh und Davis (2000) aufgestellte Konzeption des TAM2 hingegen betrachten die subjektive Norm als wichtigen sozialen Einflussfaktor, der u. a. positiv auf die Nutzenwahrnehmung und unter bestimmten Voraussetzungen auch direkt auf die Nutzungsintention wirken kann. Auch wenn andere Arbeiten keinen solchen Zusammenhang identifizieren konnten (bspw. Chismar & Wiley-Patton, 2003), könnte eine subjektive Norm, übertragen auf den Kontext von Vergleichsarbeiten, durchaus eine Rolle bei deren Wahrnehmung und Bewertung spielen.

Subjektive Norm im Sinne einer Einflussnahme des sozialen Umfeldes findet sich in diesem Zusammenhang besonders im Einfluss und den Erwartungen des kollegialen Umfeldes und der Schulleitung wieder.

Zur Rolle der Schulleitung und etwas allgemeiner dem Einfluss der Schulkultur auf die Nutzung von Vergleichsarbeiten bzw. generell die Nutzung externer Datenquellen gibt es bereits verschiedene Befunde in der Literatur. Es zeigen sich Hinweise, dass Schulleitungen wichtige Akteur*innen im Hinblick auf das Nutzungsverhalten von Lehrkräften hinsichtlich Vergleichsarbeiten und zum Anstoß einer datenbasierten Schul- und Unterrichtsentwicklung darstellen (siehe bspw. Kronsfoth et al., 2018; Wurster, Bach et al., 2016). Hierbei spielen vor allem Führungsstil und Schulkultur eine entscheidende Rolle. Datenorientierung und offene Schulkultur im Sinne eines positiven Innovations- und Kooperationsklimas erweisen sich als tendenziell förderlich für die Evidenzorientierung an Schulen und die Arbeit mit entsprechenden Daten (siehe bspw. Demski, 2016; Ercan et al., 2021; Maier et al., 2012; Zlatkin-Troitschanskaia et al., 2016). Somit liegt es auf der Hand, dass der Einfluss von Faktoren wie Schulleitungshandeln und Schulkultur in zukünftigen Untersuchungen berücksichtigt werden sollte.

Ein erneuter Blick auf das Konstrukt subjektive Norm im TAM2 verdeutlicht jedoch, dass zwar durchaus eine gewisse konzeptionelle Nähe zwischen diesem und den hier ausgeführten möglichen schulischen Einflussfaktoren im VERA-Kontext besteht, diese sich jedoch keinesfalls eins zu eins entsprechen. Subjektive Norm wäre die Wahrnehmung einer Erwartung der Schulleitung, dass Lehrkräfte mit Evaluationsdaten arbeiten. Schulkultur und Schulleitungshandeln stellen eher externe Faktoren oder Kontextfaktoren dar. Während die subjektive Norm dem herkömmlichen Verständnis nach einem Wahrnehmungskonstrukt entspricht und somit im inneren Akzeptanzmodell zu verorten wäre, werden Einfluss von Schulkultur und Schulleitung in der Modellerweiterung in Abbildung 25 als externe Faktoren positioniert, deren Einfluss auf einzelne Konstrukte des Verarbeitungsprozesses zukünftig zu untersuchen wäre.

Weitere Faktoren, die im TAM2 untersucht werden und auch für die Akzeptanz von Vergleichsarbeiten wichtig sein könnten, sind die Konstrukte Ergebnisqualität (*output quality*) und Arbeitsrelevanz (*job relevance*) (Venkatesh & Davis, 2000). Arbeitsrelevanz bezieht sich auf die Wahrnehmung einer Person, inwiefern das betrachtete System für ihre oder seine

Arbeit anwendbar bzw. bedeutsam ist. Ergebnisqualität bezieht sich auf die Wahrnehmung, wie gut ein System die Aufgaben erfüllt, die es erfüllen soll (Chismar & Wiley-Patton, 2003; Venkatesh & Davis, 2000). Für beide Faktoren finden sich im Modell Hinweise einer Beeinflussung der Nutzenwahrnehmung (siehe bspw. F. D. Davis et al., 1992; R. Davis & Wong, 2007; Venkatesh & Davis, 2000). Übertragen auf die Wahrnehmung von Vergleichsarbeiten kann die Arbeitsrelevanz als eine Art allgemeinere Nutzenbeurteilung interpretiert werden, im Sinne einer Beantwortung einer Frage nach der Bedeutung von VERA für die Arbeit von Lehrkräften im Allgemeinen. Die Bewertung der Outputqualität spielt eher in die Beurteilung der Rückmeldungen hinein, ggf. könnte es hier sinnvoll sein, diese in einem separaten Konstrukt differenziert zu erfassen.

Mit Blick auf die Beurteilung der Rückmeldungen ist neben den inhaltlichen Erwartungen, die bereits in Kapitel 6.3 angeschnitten wurden, u. a. die Verständlichkeit der Darstellungen von Bedeutung, welche im Kontext des Konstrukts zeitliche Belastung bzw. wahrgenommene Einfachheit in Kapitel 6.1.4 angesprochen wurden. Jedoch bezieht sich Verständlichkeit bzw. Verständnis im VERA-Kontext nicht zwingend nur auf die rückgemeldeten Daten, sondern auf sämtliche zur Verfügung gestellten Informationsmaterialien und den Prozess im Ganzen. Es zeigen sich Hinweise darauf, dass Qualität und Verständlichkeit der zur Verfügung stehenden Materialien das Umsetzungshandeln von Lehrkräften positiv beeinflussen können (Zuber, 2019). Daher könnten sowohl eine Beurteilung der Rückmeldungen einschließlich einer Bewertung der Verständlichkeit als Repräsentation von Outputqualität als auch ggf. eine Beurteilung der Verständlichkeit des gesamten Verfahrens in künftigen Untersuchungen des Akzeptanzmodells berücksichtigt werden.

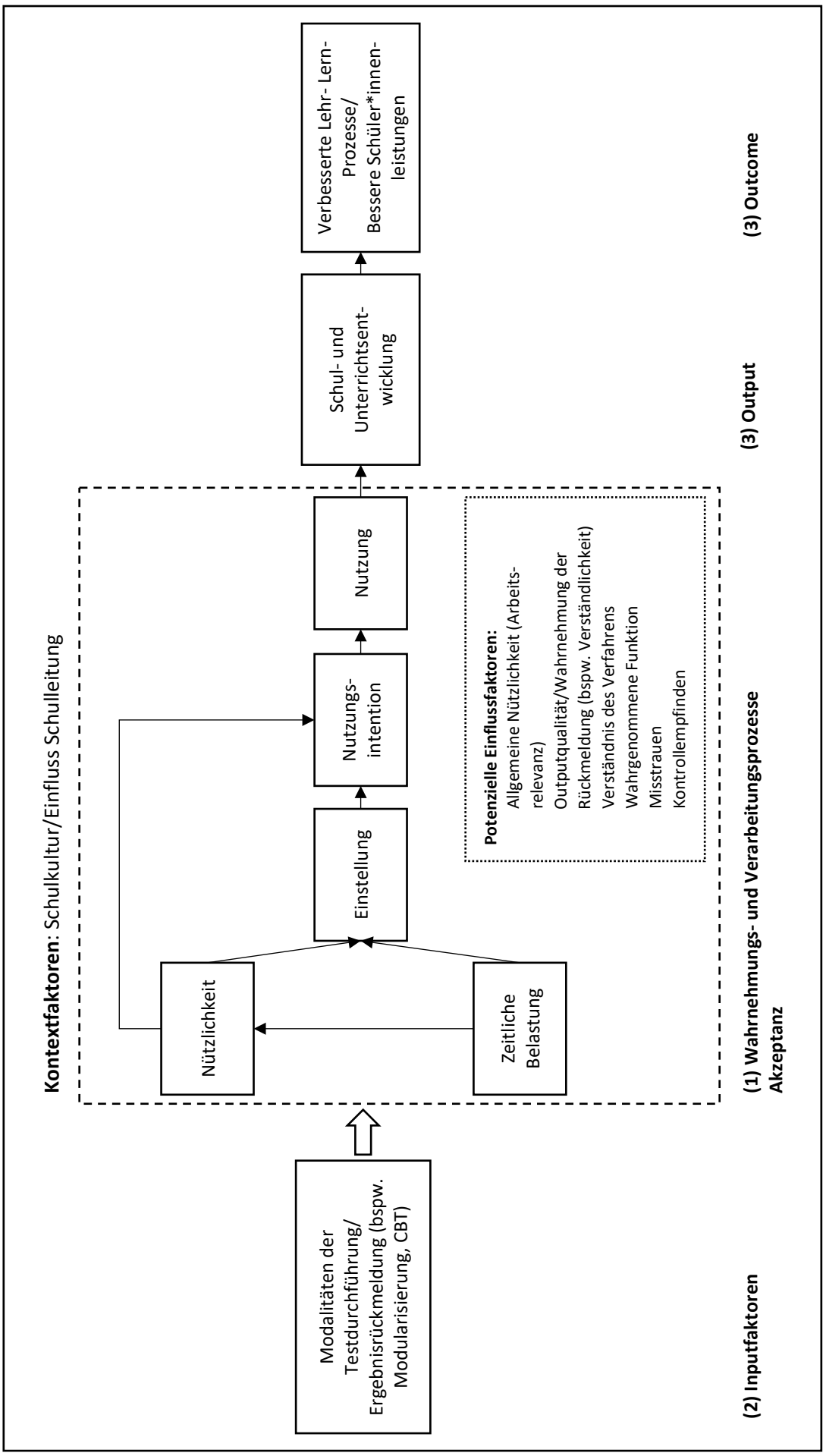


Abbildung 25: Erweiterung des Forschungsmodells

Weitere Aspekte, die in dem in dieser Arbeit untersuchten Modell noch keine Berücksichtigung fanden, deren Relevanz jedoch bereits in Kapitel 2.2.2.6 angesprochen wurde, betreffen die wahrgenommene Funktion von Vergleichsarbeiten sowie Kontrollempfinden und Misstrauen im Hinblick auf das Instrument. So könnte die Wahrnehmung der Vergleichsarbeiten als Kontrollinstrument, das Empfinden des Kontrolliertwerdens bei Lehrkräften fördern und dadurch zu unerwünschten Konsequenzen wie einer verstärken Testvorbereitung (Teaching to the Test), unerlaubten Hilfestellungen oder nicht vorgabenkonformem Auswerten der Aufgaben führen. Die Vermutung, dass Testergebnisse durch derartige Verhaltensweisen von Kolleg*innen verfälscht werden, könnte sich auf die Bewertung von Vergleichsarbeiten auswirken. Diese möglichen Zusammenhänge sollten in weiteren Untersuchungen berücksichtigt werden.

Die erläuterten Aspekte einer möglichen Modellerweiterung (Teil 1) bilden sicherlich nicht alle Faktoren ab, die für die Bewertung und Akzeptanz von Vergleichsarbeiten von Bedeutung sind, sind jedoch ein erster Schritt, künftig ein noch umfassenderes Verständnis von Akzeptanz zu erlangen. Die konkrete Verortung der Konstrukte im Modell muss bei einer Umsetzung der erweiterten Modellkonzeption eruiert werden, die Darstellung in Abbildung 25 (siehe innerer Kasten: Potenzielle Einflussfaktoren) schematische Visualisierung und repräsentiert noch keine konkreten Forschungshypothesen zu Wirkungszusammenhängen.

Der zweite Teil möglicher Modellerweiterungen setzt bei der Frage an, wie die Akzeptanz von Lehrkräften und somit die Nutzung von VERA im besten Falle zum Positiven verbessert werden kann, wie also die verschiedenen Wahrnehmungsfaktoren beeinflusst werden können. Hierzu gibt Kapitel 6.3 schon einige Hinweise, wie möglicherweise u. a. durch eine Veränderung der Test- und Durchführungsmodalitäten die Wahrnehmung von Lehrkräften verbessert werden und dadurch die Akzeptanz für das Instrument erhöht werden könnte. Die tatsächliche Auswirkung der beschriebenen Möglichkeiten der Modularisierung und eines durch eine computerbasierte Testung angepassten Testmodus und weitere daraus erwachsende Entwicklungsmöglichkeiten müssen jedoch im Weiteren noch empirisch überprüft werden. Hier würde sich bspw. ein experimentelles Design anbieten, indem die Akzeptanz der Lehrkräfte unter verschiedenen Testbedingungen (CBT vs. PP; verschiedene Stufen modularer Testheftzusammensetzung bzw. adaptiver Testheftauswahl) erhoben wird. Ein derartiges Forschungsdesign könnte Aufschluss über die Wirksamkeit verschiedener Weiterentwicklungsmaßnahmen im Blick auf die Akzeptanz von Vergleichsarbeiten liefern und könnte neben einer Ausweitung des Akzeptanzmodells zukünftig verfolgt werden.

Abschließend verbleibt die Frage nach den Anschlusshandlungen. Selbst mit dem Vorhandensein von Akzeptanz und einer Nutzung von Ergebnisrückmeldungen ist das eigentliche Ziel von Vergleichsarbeiten, eine langfristige Schul- und Unterrichtsentwicklung zu unterstützen, nicht automatisch erreicht. Dieses messbar zu machen, steht im Fokus des dritten Teils des vorgeschlagenen Forschungskonzeptes. Abbildung 25 unterscheidet hierbei zwischen dem Output von Vergleichsarbeiten, der bereits angesprochenen Schul- und Unterrichtsentwicklung, und dem daraus resultierenden Outcome, perspektivisch verbesserten Lehr-Lern-Prozessen, die sich auch in verbesserten Schüler*innenleistungen niederschlagen.

Zur Einordung in einen theoretischen Rahmen dieser dargestellten Konzeption zur Evaluation der Wirkung von Vergleichsarbeiten auf Unterrichtsqualität eignet sich das Evaluationsmodell von Kirkpatrick (1959, zitiert nach Gollwitzer & Jäger, 2014; Kirkpatrick & Kirkpatrick, 2006). Dieses Evaluationsmodell, entwickelt zur Evaluierung des Erfolgs von Trainingskonzepten, beschreibt vier Stufen der Ergebnisevaluation. Stufe 1 bezieht sich auf die direkte Reaktion der Teilnehmenden im Sinne einer unmittelbaren Bewertung. Stufe 2 beschreibt die Evaluation des Lernerfolgs. Diese Stufe umfasst u. a. auch die Erfassung einer möglichen Einstellungsänderung sowie die Motivation und Fähigkeit das Erlernte in der Praxis anzuwenden. Auf der dritten Stufe sollen das Verhalten bzw. die etwaige Verhaltensänderung infolge des Trainings evaluiert werden. Auf der vierten und letzten Stufe wird der Frage nachgegangen, inwieweit die durch eine Maßnahme angestrebten Ziele und Ergebnisse auf organisationaler Ebene erreicht wurden. Übertragen auf das hier aufgestellte Evaluationskonzept für die Nutzung von Vergleichsarbeiten (siehe Abbildung 25) lassen sich verschiedene Parallelen, insbesondere mit Blick auf Stufe 3 und 4, ziehen.

Die in Abbildung 25 unter (2) gefassten Inputfaktoren, also die variabel von außen beeinflussbaren Aspekte von Vergleichsarbeiten, bei denen ein Handlungsspielraum zur Einflussnahme auf den Charakter der Vergleichsarbeiten besteht, entsprechen hierbei dem Trainingskonzept in Kirkpatricks Modell. Ebene 1 und 2, spiegeln mit der Evaluation von Reaktion und Lernen, in den Grundzügen das in der vorliegenden Arbeit erarbeitete Akzeptanzkonzept wieder, wobei die Abgrenzung zu Ebene 3, dem Verhalten nicht eindeutig ist. Die Reaktion auf Ebene 1 ist hierbei vergleichbar mit der im Akzeptanzmodell verorteten Bewertung von Nützlichkeit und zeitlicher Belastung. Der auf Ebene 2 angesiedelte Lernerfolg könnte im Fall des

Akzeptanzmodells durch die Faktoren Einstellung bzw. Einstellungsänderung und die Motivation und ggf. auch die Intention zur Weiterarbeit mit Vergleichsarbeiten repräsentiert werden.

In Kirkpatricks Modell ist das Verhalten die Nutzung bzw. die Anwendung des Gelernten, in dem hier genutzten Akzeptanzmodell ist die Nutzung als Teil des Akzeptanzkonzepts noch vorgelagert bzw. auf einer anderen Ebene angesiedelt. Die Verhaltensänderung, Ebene 3 in Kirkpatricks Modell, wäre demnach eher auf eine mittel- bis längerfristige Veränderung, Anpassung und Verbesserung von Unterricht, also den Output von Vergleichsarbeiten, zu übertragen, nicht auf eine einmalige Auseinandersetzung mit und ggf. Nutzung von Ergebnissen, dem letzten Schritt des Akzeptanzmodells. Ebene 4, dem Erreichungsgrad der angestrebten Ergebnisse entspricht dem gewünschten langfristigen Outcome von Vergleichsarbeiten, insgesamt optimierten Lehr-Lern-Prozessen und damit einhergehenden verbesserten Leistungen der Schüler*innen.

Die Untersuchung dieses Outcomes, ebenso wie des Outputs von Vergleichsarbeiten eröffnet nun in Erweiterung der Erkenntnisse der vorliegenden Arbeit eine sehr komplexe weitere Fragestellung, die ein ebenso komplexes Forschungsdesign verlangt, mit verschiedenen zuvor zu lösenden Problemen. Dieses kann an dieser Stelle zum Abschluss dieser Arbeit nur knapp angedacht werden:

Mit Blick auf eine Evaluation des Outputs, also einer potenziellen Veränderung des Verhaltens im Sinne einer Unterrichtsentwicklung infolge der Auseinandersetzung mit und Nutzung von Rückmeldungen aus Vergleichsarbeiten, stellt sich zunächst die Frage, wie Schul- und Unterrichtsentwicklung messbar gemacht werden können. Hierbei würde sich bspw. eine qualitative Befragung von Lehrkräften zu konkreten Maßnahmen und Veränderungen infolge der Arbeit mit VERA-Ergebnissen, ggf. in Verbindung mit Unterrichtsbeobachtungen, eignen. Die Evaluation des Outcomes, Untersuchung der Wirksamkeit dieser Unterrichtsveränderung durch neu etablierte Maßnahmen wäre entsprechend, analog zu Kirkpatricks Modell die nächste Evaluationsstufe. Hier wäre ein Kontrollgruppendesign, bspw. im Rahmen eines quasi-experimentellen Designs, wünschenswert, mit Hilfe dessen die Wirksamkeit einzelner Maßnahmen beurteilt werden kann. Ist ein Kontrollgruppendesign nicht möglich, eignet sich auch eine Kohortenanalyse zur Evaluation der Wirksamkeit von Unterrichtsmaßnahmen. Angenommen es ist das Ziel, den Erfolg eines basierend auf VERA3-Ergebnissen eingeführten Förderkonzepts zu evaluieren, könnten die Leistungen des Jahrgangs, in dem das Konzept zum ersten Mal eingesetzt

wurde mit den Leistungen früherer Kohorten verglichen werden. Die forschungsleitende Fragestellung wäre demnach: Schneiden die Schüler*innen der betrachteten Kohorte zum Ende der dritten oder auch vierten Klasse besser ab, als vorangegangene Jahrgänge. Selbstverständlich stellt sich hierbei die Frage, ob und wie eine festgestellte potenzielle Verbesserung überhaupt auf die Arbeit mit VERA zurückgeführt und andere Einflussfaktoren ausgeschlossen werden können.

Generell sind hinsichtlich möglicher Untersuchungen zum Output und Outcome von Vergleichsarbeiten insgesamt noch viele methodische und konzeptionelle Überlegungen notwendig. Dennoch sollten sich zukünftige Forschungsarbeiten dieser Aufgabe widmen, u. a. um empirische Belege für den objektiven Nutzen von Vergleichsarbeiten vorzulegen. Ziel sollte sein, die praktische Relevanz von Vergleichsarbeiten zu demonstrieren und empirisch zu untermauern, um dadurch die Nutzung von VERA in der Schulpraxis zu fördern.

Literaturverzeichnis

AAPOR. (2016). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys* (9. Aufl.). Verfügbar unter: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Ackeren, I. v. & Bellenberg, G. (2004). Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen. Bestandsaufnahme und Perspektiven. In H. G. Holtappels, K. Klemm, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Bd. 13, S. 125–159). Weinheim: Juventa Verl.

Agarwal, J. & Malhotra, N. K. (2005). An Integrated Model of Attitude and Affect. *Journal of Business Research*, *58*(4), 483–493. https://doi.org/10.1016/S0148-2963(03)00138-3

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. In J. Kuhl & J. Beckmann (Hrsg.), *Action Control: From Cognition to Behavior* (S. 11–39). Berlin: Springer.

Ajzen, I. (1989). Attitude Structure and Behavior. In A. R. Pratkanis, S. J. Breckler & A. G. Greenwald (Eds.), *Attitude Structure and Function* (Ohio State university volume on attitudes and persuasion, vol. 3, S. 241–274). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I. (2002). Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior. *Journal of Applied Social Psychology*, *32*(4), 665–683. https://doi.org/10.1111/j.1559-1816.2002.tb00236.x

Ajzen, I. (2014). The Theory of Planned Behaviour is Alive and Well, and Not Ready to Retire: A Commentary on Sniehotta, Presseau, and Araújo-Soares. *Health Psychology Review*, *9*(2), 131–137. https://doi.org/10.1080/17437199.2014.883474

Ajzen, I. & Albarracin, D. (2007). Predicting and Changing Behavior: A Reasoned Action Approach. In I. Ajzen, D. Albarracin & R. C. Hornik (Hrsg.), *Prediction and change of health behavior. Applying the reasoned action approach* (S. 3–21). Mahwah, N.J.: L. Erlbaum Associates.

Ajzen, I. & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, N.J: Prentice-Hall.

Ajzen, I. & Fishbein, M. (2000). Attitudes and the Attitude-Behavior Relation: Reasoned and Automatic Processes. *European Review of Social Psychology*, *11*(1), 1–33. https://doi.org/10.1080/14792779943000116

Ajzen, I. & Fishbein, M. (2005). The Influence of Attitudes on Behavior. In D. Albarracin, B. T. Johnson & M. P. Zanna (Eds.), *The Handbook of Attitudes* (S. 173–221). Mahwah, NJ: Lawrence Erlbaum Associates.

Ajzen, I. & Madden, T. J. (1986). Prediction of Goal-Directed Behavior: Attitudes, Intentions, and Perceived Behavioral Control. *Journal of Experimental Social Psychology*, *22*(5), 453–474. https://doi.org/10.1016/0022-1031(86)90045-4

Akaike, H. (1974). A New Look on Statistical Model Identification. *IEEE Transactions on Autmatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Allport, G. W. (1935). Attitudes. In C. Murchison (Hrsg.), *Handbook of Social Psychology* (2. Aufl., S. 798–844). Worchester, MA: Clark University Press.

Allport, G. W. (1967). Attitudes. In M. Fishbein (Hrsg.), *Readings in Attitude Theory and Measurement.* New York: John Wiley & Sons.

Altrichter, H., Moosbrugger, R. & Zuber, J. (2016). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter, K. Maag Merki, R. Moosbrugger & J. Zuber (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (Educational Governance, 7 // 34, S. 235–277). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH, Wiesbaden.

Amrein, A. L. & Berliner, D. C. (2002). High-Stakes Testing & Student Learning. *education policy analysis archives*, *10*(18), 18. https://doi.org/10.14507/epaa.v10n18.2002

Anderson, J. C. & Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, *103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Antoniou, A.-S., Ploumpi, A. & Ntalla, M. (2013). Occupational Stress and Professional Burnout in Teachers of Primary and Secondary Education: The Role of Coping Strategies. *Psychology*, *04*(03), 349–355. https://doi.org/10.4236/psych.2013.43A051

Artelt, C., Schneider, W. & Schiefele, U. (2002). Ländervergleich zur Lesekompetenz. In C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 55–94). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-663-11042-2_3

Bach, A., Wurster, S., Thillmann, K., Pant, H. A. & Thiel, F. (2014). Vergleichsarbeiten und schulische Personalentwicklung – Ausmaß und Voraussetzungen der Datennutzung. *Zeitschrift für Erziehungswissenschaft*, *17*(1), 61–84. https://doi.org/10.1007/s11618-014-0486-5

Bachman, J. G. & O'Malley, P. M. (1984). Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles. *Public Opinion Quarterly*, *48*(2), 491–509. https://doi.org/10.1086/268845

Bagozzi, R. P. (1991). Further Thoughts on the Widity of Measures of Elation, Gladness, and Joy. *Journal of Personality and Social Psychology*, *61*(1), 98–104. https://doi.org/10.1037/0022-3514.61.1.98

Bagozzi, R. P. & Burnkrant, R. E. (1979). Attitude Organization and the Attitude-Behavior Relationship. *Journal of Personality and Social Psychology*, *37*(6), 913–929. https://doi.org/10.1037/0022-3514.37.6.913

Bandura, A. (1977). Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1982). Self-Efficacy Mechanism in Human Agency. *American Psychologist*, *37*(2), 122–147. https://doi.org/10.1037/0003-066X.37.2.122

Bandura, A. (1991). Social Cognitive Theory of Self-Regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

Beaujean, A. A. (2014). *Latent variable modeling using R. A step-by-step guide*. New York, London: ROUTLEDGE.

Bellmann, J., Schweizer, S. & Thiel, C. (2016). Nebenfolgen Neuer Steuerung unter Bedingungen von „low-stakes“ und „no-stakes“ – Qualitative und quantitative Befunde einer Untersuchung in vier Bundesländern. In Bundesministerium für Bildung und Forschung (Hrsg.), *Steuerung im Bildungssystem. Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (Bildungsforschung, Bd. 43, S. 208–237).

Bellmann, J. & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem. Theoretische Konzeptualisierung und Erklärungsmodelle. *Zeitschrift für Pädagogik*, *55*(2), 286–308. Zeitschrift für Pädagogik. https://doi.org/10.25656/01:4251

Bem, D. J. (1972). Self-Perception Theory. *Advances in Experimental Social Psychology*, *6*, 1–62. https://doi.org/10.1016/S0065-2601(08)60024-6

Bentler, P. M. (1990). Comparative Fit Indices in Structural Models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Berger, S. & Moser, U. (2020). Adaptives Lernen und Testen. *journal für lehrerInnenbildung*, *20*(1), 42–52. https://doi.org/10.35468/jlb-01-2020_03

Berliner, D. C. (2011). Rational Responses to High Stakes Testing: The Case of Curriculum Narrowing and the Harm that Follows. *Cambridge Journal of Education*, *41*(3), 287–302. https://doi.org/10.1080/0305764X.2011.607151

Berning, C. C. (2019). Strukturgleichungsmodelle. In M. Apelt, I. Bode, R. Hasse, U. Meyer, V. V. Groddeck, M. Wilkesmann et al. (Hrsg.), *Handbuch Organisationssoziologie*

(Springer Reference Sozialwissenschaften, Bd. 38, S. 1–18). Wiesbaden: Springer Fachmedien Wiesbaden; Springer VS. https://doi.org/10.1007/978-3-658-16937-4_30-2

Bez, S., Poindl, S., Bohl, T. & Merk, S. (2021). Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien. *Zeitschrift für Pädagogik*, *67*(4), 552-572. https://doi.org/10.3262/ZP2104551

Bohner, G. & Schwarz, N. (2001). Attitudes, Persuasion, and Behavior. In A. Tesser & N. Schwarz (Hrsg.), *Blackwell handbook of social psychology. Intraindividual processes* (S. 413–435). Oxford, UK: Blackwell.

Bollen, K. A. (2000). Modeling Strategies: In Search of the Holy Grail. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(1), 74–81. https://doi.org/10.1207/S15328007SEM0701_03

Bollen, K. A. (2014). *Structural Equations with Latent Variables* (Wiley Series in Probability and Statistics). Hoboken: Wiley.

Bonsen, M., Büchter, A. & Peek, R. (2006). Datengestützte Schul- und Unterrichtsentwicklung. Bewertung der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Bd. 14, 1. Aufl., Bd. 14, S. 125–148). Weinheim: Juventa Verl.

Boomsma, A. (2000). Reporting Analyses of Covariance Structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(3), 461–483. https://doi.org/10.1207/S15328007SEM0703_6

Bosnjak, M. & Tuten, T. L. (2001). Classifying Response Behaviors in Web-based Surveys. *Journal of Computer-Mediated Communication*, *6*(3), 0. https://doi.org/10.1111/j.1083-6101.2001.tb00124.x

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *53*, 345–370.

Breckler, S. J. (1984). Empirical Validation of Affect, Behavior, and Cognition as Distinct Components of Attitude. *Journal of Personality and Social Psychology*, *47*(6), 1191–1205. https://doi.org/10.1037/0022-3514.47.6.1191

Brosseau-Liard, P. E. & Savalei, V. (2014). Adjusting Incremental Fit Indices for Nonnormality. *Multivariate behavioral research*, *49*(5), 460–470. https://doi.org/10.1080/00273171.2014.933697

Brosseau-Liard, P. E., Savalei, V. & Li, L. (2012). An Investigation of the Sample Performance of Two Nonnormality Corrections for RMSEA. *Multivariate behavioral research*, *47*(6), 904–930. https://doi.org/10.1080/00273171.2012.715252

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (Methodology in the social sciences, Second edition). New York, London: The Guilford Press.

Brown, T. A. & Moore, M. T. (2012). Confirmatory Factor Analysis. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 361–379). New York [u.a.]: Guilford Press.

Browne, M. W. & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Campbell, D. T. & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Chaiken, S., Libermann, A. & Eagly, A. H. (1989). Heuristic and Systematic Information Processing within and beyond the Persuasion Context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (S. 212–252). New York: Guilford Press.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Chismar, W. G. & Wiley-Patton, S. (2003). Does the Extended Technology Acceptance Model Apply to Physicians. In R. H. Sprague (ed.), *Proceedings of the 36th Annual Hawaii International Conference on System Sciences physicians* (8 pp). Los Alamitos, Calif: IEEE Computer Society Press.

Christophersen, T. & Grape, C. (2009). Die Erfassung latenter Konstrukte mit Hilfe formativer und reflektiver Messmodelle. In S. Albers, D. Klapper, U. Konradt, J. Wolf & A. Walter (Hrsg.), *Methodik der empirischen Forschung* (3., überarbeitete und erweiterte Auflage, S. 103–118). Wiesbaden: Gabler Verlag.

Cizek, G. J. (2001). More Unintended Consequences of High-Stakes Testing. *Educational Measurement: Issues and Practice*, *20*(4), 19–27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Aufl.). New York: Erlbaum.

Couper, M. P. (2000). Review: Web Surveys: A Review of Issues and Approaches. *The Public Opinion Quarterly*, *64*(4), 464–494.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297–334.

Davis, F. D. (1986). *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*. Dissertation. Sloan School of Management.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Davis, F. D. (1993). User Acceptance of Information Technology: System Characteristiccs, User Perceptions and Behavioral Impacts. *International Journal of Man-Machine Studies*, *38*(3), 475–487. https://doi.org/10.1006/imms.1993.1022

Davis, F. D., Bagozzi, R. P. & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science*, *35*(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

Davis, F. D., Bagozzi, R. P. & Warshaw, P. R. (1992). Extrinsic and Intrinsic Motivation to Use Computers in the Workplace. *Journal of Applied Social Psychology*, *22*(14), 1111–1132. https://doi.org/10.1111/j.1559-1816.1992.tb00945.x

Davis, R. & Wong, D. (2007). Conceptualizing and Measuring the Optimal Experience of the eLearning Environment. *Decision Sciences Journal of Innovative Education*, *5*(1), 97–126. https://doi.org/10.1111/j.1540-4609.2007.00129.x

Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen. *Unterrichtswissenschaft*, *39*(1), 63–83.

Demski, D. (2016). *Evidenzbasierte Schulentwicklung. Empirische Analyse eines Steuerungsparadigmas* (Schulentwicklungsforschung, Bd. 2). Dissertation. Universität Duisburg-Essen. https://doi.org/10.1007/978-3-658-18078-2

Demski, D. (2019a). Nutzung von internen und externen Evaluationen in der Schulpraxis. In T. Stricker (Hrsg.), *Zehn Jahre Fremdevaluation in Baden-Württemberg. Zwischenbilanz und Perspektiven auf Qualitätsmanagement, Evaluation und Schulentwicklung* (S. 45–56). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-25778-1_4

Demski, D. (2019b). Und was kommt in der Praxis an? Bewertung und Nutzung von Instrumenten der Neuen Steuerung durch Schulleitungsmitglieder und Lehrkräfte // Und was kommt in der Praxis an? In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (Educational Governance, Bd. 42, S. 129–152). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-22241-3_6

Diemer, T. (2013). *Innerschulische Wirklichkeiten neuer Steuerung. Zur Nutzung zentraler Lernstandserhebungen* (1. Aufl.). Dissertation. Wiesbaden: Springer VS.

Diemer, T., Hartung-Beck, V. & Kuper, H. (2013). Die Abnehmerperspektive: Rückmeldeforschung im Kontext schulischer Evaluation mittels zentraler Lernstandserhebungen. In M. Rürup & I. Bormann (Hrsg.), *Innovationen im Bildungswesen. Analytische Zugänge und empirische Befunde* (Educational Governance, Bd. 21, S. 173–188). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-531-19701-2_8

Diemer, T. & Kuper, H. (2011). Formen innerschulischer Steuerung mittels zentraler Lernstandserhebungen. *Zeitschrift für Pädagogik*, *57*(4), 554–571. Zeitschrift für Pädagogik 57 (2011) 4, S. 554-571. https://doi.org/10.25656/01:8746

Dillon, W. R. & Kumar, A. [Ajith]. (1985). Attitude Organization and the Attitude-Behavior Relation: A Critique of Bagozzi and Burnkrant's Reanalysis of Fishbein and Ajzen. *Journal of Personality and Social Psychology*, *49*(1), 33–46. https://doi.org/10.1037/0022-3514.49.1.33

Ditton, H., Merz, D. & Edelhäußer, T. (2002). Einstellungen von Lehrkräften und Schulleiter/innen zu zentralen Testuntersuchungen an Schulen. *Empirische Pädagogik*, *16*(1), 17–33.

Eagly, A. H. & Chaiken, S. (1993). *The Psychology of Attitudes*. Forth Worth: TX: Harcourt, Brace, & Janovich.

Eagly, A. H. & Chaiken, S. (1998). Attitude Structure and Function. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Hrsg.), *The Handbook of Social Psychology, Vols. 1 and 2 (4th ed.)* (S. 269–322). New York: McGraw-Hill.

Eagly, A. H. & Chaiken, S. (2005). Attitude Research in the 21st Century. The Current State of Knowledge. In D. Albarracin, B. T. Johnson & M. P. Zanna (Eds.), *The Handbook of Attitudes* (S. 743–767). Mahwah, NJ: Lawrence Erlbaum Associates.

Eberl, M. (2004). *Formative und reflektive Indikatoren im Forschungsprozess: Entscheidungsregeln und die Dominanz des reflektiven Modells* (Ludwig-Maximilians-Universität München: Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung 19).

Eckardt, G. (2015). *Sozialpsychologie - Quellen zu ihrer Entstehung und Entwicklung* (Schlüsseltexte der Psychologie). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-06854-7

Edwards, J. R. (2011). The Fallacy of Formative Measurement. *Organizational Research Methods*, *14*(2), 370–388. https://doi.org/10.1177/1094428110378369

Edwards, J. R. & Bagozzi, R. P. (2000). On the Nature of Direction of Relationships Between Constructs and Measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Emons, W. H. M. (2008). Nonparametric Person-Fit Analysis of Polytomous Item Scores. *Applied Psychological Measurement*, *32*(3), 224–247. https://doi.org/10.1177/0146621607302479

Enders, C. K. (2001). The Impact of Nonnormality on Full Information Maximum-Likelihood Estimation for Structural Equation Models With Missing Data. *Psychological methods*, *6*(4), 352–370. https://doi.org/10.1037/1082-989X.6.4.352

Enders, C. K. (2010). *Applied Missing Data Analysis* (Methodology in the social sciences). New York: Guilford Press.

Enders, C. K. & Bandalos, D. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Engel, U. & Schmidt, B. O. (2014). Unit- und Item-Nonresponse. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 331–348). Wiesbaden: Springer VS.

Ercan, H., Hartmann, U., Richter, D., Kuschel, J. & Gräsel, C. (2021). Effekte von integrativer Führung auf die Datennutzung von Lehrkräften. *DDS – Die Deutsche Schule*, *2021*(1), 85–100. https://doi.org/10.31244/dds.2021.01.08

Fan, X., Thompson, B. & Wang, L. (1999). Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56–83. https://doi.org/10.1080/10705519909540119

Fazio, R. H. (1986). How do Attitudes Guide Behavior? In R. M. Sorrentino & E. T. Higgins (Hrsg.), *Handbook of motivation and cognition: Foundations of social behavior* (S. 204–243). New York: Guilford Press.

Felser, G. (2015). *Werbe- und Konsumentenpsychologie* (4. Aufl.). Berlin: Springer. https://doi.org/10.1007/978-3-642-37645-0

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford: Stanford Univ. Press.

Festinger, L. (1964). Behavioral Support for Opinion Change. *Public Opinion Quarterly*, *28*(3), 404–417. https://doi.org/10.1086/267263

Fiege, C., Reuther, F. & Nachtigall, C. (2011). Faire Vergleiche? – Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen

Vergleichsarbeiten. *Zeitschrift für Bildungsforschung*, *1*(2), 133–149. https://doi.org/10.1007/s35834-011-0009-x

Finch, W. H. & French, B. F. (2015). *Latent Variable Modeling with R*. New York: Routledge Taylor & Francis Group.

Fischer, P., Asal, K. & Krueger, J. I. (Hrsg.). (2014). *Sozialpsychologie für Bachelor. Lesen, Hören, Lernen im Web* (Bachelor). Berlin: Springer.

Fishbein, M. & Ajzen, I. (1975). *Belief, Attitude, Intention and Behavior. An Introduction to Theory and Research* (Addison-Wesley series in social psychology). Reading, Mass.: Addison-Wesley.

Fishbein, M. & Ajzen, I. (2010). *Predicting and Changing Behavior. The Reasoned Action Approach*. New York: Psychology Press.

Fluck, J. (2020a). *Cyberbullying. Theoretische und empirische Analysen zur Konstruktklärung und Messmodellierung eines Gewaltphänomens*. Hamburg: Verlag Dr. Kovač.

Fluck, J. (2020b). *Formative Messmodelle und Möglichkeiten ihrer Anwendung im empirisch-pädagogischen Kontext. Datengeleitete Indexbildung nach der MARI-Methode*: RWTH Aachen: Institut für Erziehungswissenschaft.

Fornell, C. & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Fox, J. & Weisberg, S. (2019). An R Companion to Applied Regression (Version 3.0-10) [Computer software]. Thousand Oaks CA: Sage. Verfügbar unter: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fuchs, A. (2011). *Methodische Aspekte linearer Strukturgleichungsmodelle. Ein Vergleich von kovarianz- und varianzbasierten Kausalanalyseverfahren* (Research papers on marketing strategy, Bd. 2). Würzburg: Julius-Maximilians-Universität Würzburg, Lehrstuhl für BWL und Marketing.

Fuchs, G. & Brunner, M. (2017). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie*, *31*(1), 27–39. https://doi.org/10.1024/1010-0652/a000195

Gana, K. & Broc, G. (2019). *Structural equation modeling with lavaan* (Mathematics and statistics). London, Hoboken NJ: ISTE Ltd; John Wilery & Sons Inc.

Gasteiger, H. & Krelle, M. (2018). VERA - und was dann? 10 Jahre Vergleichsarbeiten. *Grundschulmagazin*, *86*(4), 7–11.

Gathen, J. v. d. (2006). Die innerschulische Rezeption von Leistungsrückmeldungen aus Large-Scale-Assessments - Grundlagen und Ziele von Fallstudien. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 77–88). Münster: Waxmann.

Gay, L. R., Mills, G. E. & Airasian, P. (2011). *Educational Research: Competencies for Analysis and Applications* (10. Aufl.). Upper Saddle River, New Jersey: Pearson Education.

Gollwitzer, M. & Jäger, R. S. (2014). *Evaluation kompakt. Mit Arbeitsmaterial zum Download* (Kompakt, 2. Aufl.). Weinheim, Basel: Beltz.

Grabensberger, E., Freudenthaler, H. H. & Specht, W. (2008). *Bildungsstandards: Testungen und Ergebnisrückmeldungen auf der achten Schulstufe aus der Sicht der Praxis. Ergebnisse einer Befragung von Leiterinnen, Leitern und Lehrkräften der Pilotschulen* (BIFIE-Report). Graz: Bifie - Bundesinst. für Bildungsforschung, Innovation und Entwicklung des Österr. Schulwesens.

Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie*, *30*(4), 201–211. https://doi.org/10.1024/1010-0652/a000182

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing Data. Analysis and Design* (Statistics for social and behavioral sciences). New York: Springer. https://doi.org/10.1007/978-1-4614-4018-5

Graham, J. W., Cumsille, P. E. & Elek-Fisk, E. (2003). Methods for Handling Missing Data. In I. B. Weiner & J. A. Schinka (Eds.), *Handbook of Psychology. Volume 2: Research Methods in Psychology* (S. 87–114). Hoboken, NJ: Wiley.

Groß Ophoff, J. (2013). *Lernstandserhebungen: Reflexion und Nutzung*. Münster: Waxmann.

Groß Ophoff, J., Koch, U., Helmke, A. & Hosenfeld, I. (2006). Vergleichsarbeiten für die Grundschulen - und was diese daraus machen (können). *Journal für Schulentwicklung*, *10*(4), 7–12.

Groß Ophoff, J., Koch, U. & Hosenfeld, I. (2019). Vergleichsarbeiten in der Grundschule von 2004 bis 2015. Trends in der Akzeptanz und Auseinandersetzung mit Rückmeldungen. In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (Educational Governance, Bd. 42, S. 205–228). Wiesbaden: Springer VS.

Gulek, C. (2003). Preparing for High-Stakes Testing. *Theory into practice*, *42*(1), 42–50. https://doi.org/10.1207/s15430421tip4201_6

Haddock, G. & Maio, G. R. (2014). Einstellungen. In K. Jonas, W. Stroebe & M. Hewstone (Hrsg.), *Sozialpsychologie* (Springer-Lehrbuch, 6. Aufl., S. 197–229). Berlin, Heidelberg: Springer.

Haddock, G. & Maio, G. R. (2019). Inter-individual Differences in Attitude Content: Cognition, Affect, and Attitudes. In J. M. Olson (Hrsg.), *Advances in Experimental Social Psychology* (Bd. 59, S. 53–102). Amsterdam [etc.]: Elsevier. https://doi.org/10.1016/bs.aesp.2018.10.002

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2010). *Multivariate Data Analysis. A Global Perspective* (7th ed.). Upper Saddle River, NJ, Munich: Pearson.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Richter, N. F. & Hauff, S. (2017). *Partial Least Squares Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung*. München: Franz Vahlen.

Hartung-Beck, V. & Diemer, T. (2009). Sensemaking durch Outputorientierung. Erfahrungen mit der Nutzung von Lernstandserhebungen in Schulen. In M. Göhlich, S. M. Weber & S. Wolff (Hrsg.), *Organisation und Erfahrung. Beiträge der AG Organisationspädagogik* (Organisation und Pädagogik, Bd. 7, 1. Aufl., S. 215–225). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH. https://doi.org/10.1007/978-3-531-91660-6_19

Heider, F. (1946). Attitudes and Cognitive Organization. *The Journal of Psychology*, *21*, 107–112. https://doi.org/10.1080/00223980.1946.9917275

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Das Seminar*, (2), 90–112.

Helmke, A. & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In G. Brägger, B. Bucher, N. Landwehr & W. Böttcher (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127–151). Bern: Hep.

Hemmerich, W. (2015). *StatistikGuru: Cohen's d berechnen*. Verfügbar unter: https://statistikguru.de/rechner/cohens-d.html

Hermida, R. (2015). The Problem of Allowing Correlated Errors in Structural Equation Modeling: Concerns and Considerations. *Computational Methods in Social Sciences*, *2*(1), 5–17.

Ho, M. R., Stark, S. & Chernyshenko, O. (2012). Graphical Representation of Structural Equation Models Using Path Diagrams. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 43–55). New York [u.a.]: Guilford Press.

Ho Cheong, J. & Park, M.-C. (2005). Mobile Internet Acceptance in Korea. *Internet Research*, *15*(2), 125–140. https://doi.org/10.1108/10662240510590324

Hu, L. & Bentler, P. M. (1995). Evaluating Model Fit. In R. H. Hoyle (Ed.), *Structural Equation Modeling. Concepts, Issues, and Applications* (1st ed., S. 76–99). Thousand Oaks: SAGE Publications.

Hu, L. & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M. & DeShon, R. P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Isaac, K., Halt, A. C., Hosenfeld, I., Helmke, A. & Groß Ophoff, J. (2006). VERA: Qualitätsentwicklung und Lehrerprofessionalisierung durch Vergleichsarbeiten. *Die Deutsche Schule*, *98*, 107–110.

Jackson, D. L. (2003). Revisiting Sample Size and Number of Parameter Estimates: Some Support for the N:q Hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Jäger, S. (2011). *Rezeption und Nutzung von Diagnose- und Vergleichsarbeiten an Schulen. Eine Interviewstudie*. Dissertation. Verfügbar unter: http://d-nb.info/1051226406/34

Johnson, J. A. (2005). Ascertaining the Validity of Individual Protocols From Web-Based Personality Inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Jones, B. D. & Egley, R. J. (2004). Voices from the Frontlines: Teachers' Perceptions of High-Stakes Testing. *Education Policy Analysis Archives*, *12*(39), 1–34. https://doi.org/10.14507/epaa.v12n39.2004

Jones, G. M., Jones, B. D. & Hargrove, T. Y. (2003). *The Unintended Consequences of High-Stakes Testing*. Lanham: Rowman & Littlefield Publishers.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger (Hrsg.), *Testtheorie und Fragebogenkonstruktion. Mit 66 Abbildung und 41 Tabellen* (Springer-Lehrbuch, 2., aktual. und überarb. Aufl., 27-74). Berlin: Springer.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M. & Rosseel, Y. (2021). semTools: Useful Tools for Structural Equation Modeling (Version R package version 0.5-4) [Computer software]. Verfügbar unter: https://CRAN.R-project.org/package=semTools

Kaczmirek, L. (2008). *Human-Survey Interaction. Usability and Nonresponse in Online Surveys*. Dissertation. Universität Mannheim, Mannheim.

Katz, D. (1960). The Functional Approach to the Study of Attitudes. *Public Opinion Quarterly*, *24*(2), 163–204. https://doi.org/10.1086/266945

Kavita, K. & Hassan, N. C. (2018). Work Stress among Teachers: A Comparison between Primary and Secondary School Teachers. *International Journal of Academic Research in Progressive Education and Development*, *7*(4). https://doi.org/10.6007/IJARPED/v7-i4/4802

Keith, T. Z. (2019). *Multiple Regression and Beyond. An Introduction to Multiple Regression and Structural Equation Modeling* (Third edition). New York, NY: ROUTLEDGE.

Kenny, D. A. & Milan, S. (2012). Identification. A Nontechnical Discussion of a Technical Issue. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 145–163). New York [u.a.]: Guilford Press.

King, W. R. & He, J. (2006). A Meta-Analysis of the Technology Acceptance Model. *Information & Management*, *43*(6), 740–755. https://doi.org/10.1016/j.im.2006.05.003

Kirkpatrick, D. L. & Kirkpatrick, J. D. (2006). *Evaluating Training Programs. The Four Levels* (3rd ed.). San Francisco: Berrett-Koehler.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bonn, Berlin: BMBF. https://doi.org/10.25656/01:20901

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (Methodology in the social sciences, 3. Aufl.). New York: Guilford Press.

Kline, R. B. (2012). Assumptions in Structural Equation Modeling. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (111-125). New York [u.a.]: Guilford Press.

KMK. (2001). *Pressemitteilung der 280. Plenarsitzung,* Ständige Konferenz der Kultusminister der Länder in der Bundesrepubilk Deutschland. Verfügbar unter: https://www.kmk.org/presse/pressearchiv/mitteilung/297-plenarsitzung-der-kultusministerkonferenz-am-28-februar-01-maerz-2002-in-berlin.html

KMK. (2004a). *Standards für die Lehrerbildung: Bildungswissenschaften. (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 07.10.2022)*. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

KMK. (2004b). *Vereinbarung über Bildungsstandards für den Primarbereich (Jahrgangsstufe 4). Beschluss der Kultusministerkonferenz vom 15.10.2004*. Verfügbar unter:

http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Primar.pdf

KMK. (2010). *Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung*. Verfügbar unter: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2010/2010_00_00-Konzeption-Bildungsstandards.pdf

KMK. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Verfügbar unter: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

KMK. (2018). *Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (VERA). Beschluss der Kultusministerkonferenz vom 08.03.2012 i. d. F. vom 15.03.2018*. Verfügbar unter: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Koch, U. (2011). *Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten? Datenkompetenz von Lehrkräften und die Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten* (Empirische Erziehungswissenschaft 31). Münster: Waxmann.

Koch, U., Groß Ophoff, J., Hosenfeld, I. & Helmke, A. (2006). Von der Evaluation zur Schul- und Unterrichtsentwicklung - Ergebnisse der Lehrerbefragungen zur Auseinandersetzung mit den VERA-Rückmeldungen. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 187–199). Münster: Waxmann.

Koch, U. & Hosenfeld, I. (2013). Wie objektiv werden Leseverstehensaufgaben im Rahmen der Vergleichsarbeiten in der Grundschule ausgewertet? In M. Zimmer-Müller & I. Hosenfeld (Hrsg.), *Zehn Jahre Vergleichsarbeiten: eine Zwischenbilanz aus verschiedenen Perspektiven* (Empirische Pädagogik, Bd. 27,4, S. 474–496). Landau in der Pfalz: Verlag Empirische Pädagogik.

Koh, K. H. & Zumbo, B. D. (2008). Multi-Group Confirmatory Factor Analysis for Testing Measurement Invariance in Mixed Item Format Data. *Journal of Modern Applied Statistical Methods*, *7*(2), 471–477. https://doi.org/10.22237/jmasm/1225512660

Kongcharoen, J., Onmek, N., Jandang, P. & Wangyisen, S. (2020). Stress and Work Motivation of Primary and Secondary School Teachers. *Journal of Applied Research in Higher Education*, *12*(4), 709–723. https://doi.org/10.1108/JARHE-04-2019-0088

Krelle, M. (2015). Leseverstehen im Kontext der Vergleichsarbeiten für die dritte Jahrgangsstufe im Fach Deutsch – Leistungen und Grenzen eines diagnostischen Instruments zur Sprachförderung. *leseforum.ch*, (1), 1–27. Verfügbar unter: http://leseforum.ch/myUploadData/files/2015_1_Krelle.pdf

Kronsfoth, K., Muslic, B., Graf, T. & Kuper, H. (2018). Der Zusammenhang zwischen Führungsdimensionen in der Schulleitung und der Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten. *DDS - Die Deutsche Schule*, *110*(1), 47–64.

Kühle, B. & Peek, R. (2007). Lernstandserhebungen in Nordrhein-Westfalen. Evaluationsbefunde zur Rezeption und zum Umgang mit Ergebnisrückmeldungen in Schulen. *Empirische Pädagogik*, *21*(4), 428–447.

Kuhn, H.-J. (2014). Anspruch, Wirklichkeit und Perspektiven der Gesamtstrategie der KMK zum Bildungsmonitoring. *DDS - Die Deutsche Schule*, *106*(4), 414–426.

Kulviwat, S., Bruner II, G. C., Kumar, A. [Anand], Nasco, S. A. & Clark, T. (2007). Toward a Unified Theory of Consumer Acceptance Technology. *Psychology and Marketing*, *24*(12), 1059–1084. https://doi.org/10.1002/mar.20196

Kuper, H. (2008). Wissen – Evaluation – Evaluationswissen. In T. Brüsemeister & K.-D. Eubel (Hrsg.), *Evaluation, Wissen und Nichtwissen* (1. Aufl., 61-73). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuper, H. & Diemer, T. (2012). Vergleichsarbeiten: Theoretische und empirische Betrachtungen zum Nutzen des Vergleichens. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (SpringerLink : Bücher, Bd. 9, S. 225–245). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuper, H., Maier, U., Graf, T., Muslic, B. & Ramsteck, C. (2016). Datenbasierte Schulentwicklung mit Vergleichsarbeiten aus der Perspektive von Lehrkräften, Fachkonferenzleitungen, Schulleitungen und Schulaufsichten – Qualitative Fallstudien aus vier Bundesländern. In Bundesministerium für Bildung und Forschung (Hrsg.), *Steuerung im Bildungssystem. Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (Bildungsforschung, Bd. 43, S. 39–67).

Lei, P.-W. & Wu, Q. (2012). Estimation in Structural Equation Modeling. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (164-180). New York [u.a.]: Guilford Press.

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2008). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft* (Zeitschrift für Erziehungswissenschaft Sonderheft, Bd. 8, S. 149–167). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH Wiesbaden. https://doi.org/10.1007/978-3-531-90865-6_9

Levene, H. (1960). Robust Tests for Equality of Variances. In I. Olkin (Hrsg.), *Contributions to Probability and Statistics* (S. 278–292). Stanford: Stanford University Press.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *140*, 5–53.

Limesurvey GmbH. LimeSurvey: An Open Source survey tool [Computer software]. Hamburg: Limesurvey GmbH. Verfügbar unter: http://www.limesurvey.org

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119013563

Lucke, D. (1995). *Akzeptanz. Legitimität in der „Abstimmungsgesellschaft"*. Opladen: Leske + Budrich.

Ma, Q. & Liu, L. (2004). The Technology Acceptance Model: A Meta-Analysis of Empirical Findings. *Journal of Organizational and End User Computing*, *16*(1), 59–72.

Maag Merki, K. (2016). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter, K. Maag Merki, R. Moosbrugger & J. Zuber (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (Educational Governance, 7 // 34, S. 151–181). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH, Wiesbaden. https://doi.org/10.1007/978-3-531-92245-4_6

Maaz, K., Emmrich, R., Kropf, M. & Gärtner, H. (2019). Bildungsstandards als innovative Elemente moderner Bildungssysteme. In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (Educational Governance, Bd. 42, 25-44). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-22241-3_2

MacCallum, R. C. & Austin, J. T. (2000). Applications of Structural Equation Modeling in Psychological Research. *Annual Review of Psychology*, *51*(1), 201–226. https://doi.org/10.1146/annurev.psych.51.1.201

Maier, U. (2007). Welche Konsequenzen ziehen Mathematiklehrkräfte aus verpflichtenden Diagnose- und Vergleichsarbeiten? *mathematica didactica*, *30*(2), 5–32. https://doi.org/10.18716/ojs/md/2007.1083

Maier, U. (2008a). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik*, *54*(1), 95–117. https://doi.org/10.25656/01:4338

Maier, U. (2008b). Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. *Zeitschrift für Erziehungswissenschaft*, *11*(3), 453–474. https://doi.org/10.1007/s11618-008-0036-0

Maier, U. (2009a). Professionelle Nutzung von Vergleichsarbeiten? Ergebnisse einer qualitativen Interviewstudie mit Lehrkräften in Baden-Württemberg. In T. Bohl (Hrsg.), *Lernen*

aus Evaluationsergebnissen. Verbesserungen planen und implementieren (1. Aufl., S. 131–144). Bad Heilbrunn: Klinkhardt.

Maier, U. (2009b). Testen und dann? Ergebnisse einer qualitativen Lehrerbefragung zur diagnostischen Funktion von Vergleichsarbeiten. *Empirische Pädagogik*, *23*(2), 191–207.

Maier, U. (2009c). Towards state-mandated testing in Germany: how do teachers assess the pedagogical relevance of performance feedback information? *Assessment in Education: Principles, Policy & Practice*, *16*(2), 205–226. https://doi.org/10.1080/09695940903076030

Maier, U. (2009d). *Wie gehen Lehrerinnen und Lehrer mit Vergleichsarbeiten um? Eine Studie zu testbasierten Schulreformen in Baden-Württemberg und Thüringen Schul- und Unterrichtsforschung* (Schul- und Unterrichtsforschung, Bd. 7, 1. Aufl.). Schneider Verlag Hohengehren.

Maier, U. (2010a). Accountability policies and teachers' acceptance and usage of school performance feedback – a comparative study. *School Effectiveness and School Improvement*, *21*(2), 145–165. https://doi.org/10.1080/09243450903354913

Maier, U. (2010b). Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht. Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar? *Zeitschrift für Pädagogik*, *56*(1), 112–128. https://doi.org/10.25656/01:7138

Maier, U., Bohl, T., Kleinknecht, M. & Metz, K. (2011). Einflüsse von Merkmalen des Testsystems und Schulkontextfaktoren auf die Akzeptanz und Rezeption von zentralen Testrückmeldungen durch Lehrkräfte. *Journal for educational research online*, *3*(2), 62–93. https://doi.org/10.25656/01:5625

Maier, U. & Kuper, H. (2012). Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen. Überblick zum Forschungsstand. *DDS - Die Deutsche Schule*, *104*(1), 88–99. https://doi.org/10.25656/01:25723

Maier, U., Metz, K., Bohl, T., Kleinknecht, M. & Schymala, M. (2012). Vergleichsarbeiten als Instrument der datenbasierten Schul- und Unterrichtsentwicklung in Gymnasien. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (SpringerLink : Bücher, Bd. 9, S. 197–224). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maier, U. & Rauin, U. (2006). Vergleichsarbeiten - Hilfe zur Unterrichtsentwicklung? Zentrale Lernstandserhebungen aus Sicht baden-württembergischer Lehrkräfte. *Die Deutsche Schule*, *98*(4), 403–421. https://doi.org/10.25656/01:27347

Maio, G. R. & Haddock, G. (2010). *The Psychology of Attitudes and Attitude Change* (Sage social psychology program series, 1. publ // Repr). Los Angeles: Sage.

Maio, G. R. & Haddock, G. (2015). *The Psychology of Attitudes and Attitude Change* (2. ed.). Los Angeles, Calif.: Sage.

Malhotra, N. K. (2005). Attitude and Affect: New Frontiers of Research in the 21st Century. *Journal of Business Research*, *58*(4), 477–482. https://doi.org/10.1016/S0148-2963(03)00146-2

McGuire, W. J. (1968). Personality and attitude change: An information-processing theory. In A. G. Greenwald, T. C. Brock & T. M. Ostrom (Eds.), *Psychological Foundations of Attitudes* (S. 171–196). New York: Academic Press.

McGuire, W. J. (1986). The Vicissitudes of Attitudes and Similar Representational Constructs in Twentieth Century Psychology. *European Journal of Social Psychology*, *16*(2), 89–130. https://doi.org/10.1002/ejsp.2420160202

Meade, A. W. & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meade, A. W., Johnson, E. C. & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meredith, W. (1993). Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. & Olivera-Aguilar, M. (2012). Investigating Measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 380–392). New York [u.a.]: Guilford Press.

Monecke, A. & Leisch, F. (2012). semPLS: Structural Equation Modeling UsingPartial Least Squares, *Journal of Statistical Software*(48), 1–32.

Müller, P. & Schäfer, S. (2017). Latent Mean (Comparison). In J. Matthes (Ed.), *The international encyclopedia of communication research methods* (vol. 30, S. 1–7). Hoboke, New Jersey: Wiley-Blackwell. https://doi.org/10.1002/9781118901731.iecrm0132

Muslic, B. (2017). *Kopplungen und Entscheidungen in der Organisation Schule. Organisationsbezogenes Schulleitungshandeln im Kontext von Lernstandserhebungen*. Wiesbaden: Springer Fachmedien Wiesbaden.

Nayak, M. & Narayan, K. A. (2019). Strengths and Weakness of Online Surveys. *IOSR Journal of Humanities and Social Sciences*, *24*(5), 31–38.

Nichols, S. L. & Berliner, D. C. (2005). *The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing*. Tempe: Arizona State University. Verfügbar unter: https://nepc.colorado.edu/sites/default/files/EPSL-0503-101-EPRU.pdf

Niedersächsisches Kultusministerium. (2019). *Vergleichsarbeiten (VERA),* Niedersächsisches Kultusministerium. Verfügbar unter: https://www.mk.niedersachsen.de/startseite/schule/schulqualitat/externe_evaluation/vergleichsarbeiten_vera/vergleichsarbeiten-vera-135419.html

Niessen, A. S. M., Meijer, R. R. & Tendeiro, J. N. (2016). Detecting Careless Respondents in Web-Based Questionnaires: Which Method to Use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Nunnally, J. C. (1978). *Psychometric Theory* (2. Aufl.). New York: McGraw-Hill.

Nusser, L., Carstensen, C. H. & Artelt, C. (2015). Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz. *Empirische Sonderpädagogik*, *7*(2), 99–116. https://doi.org/10.25656/01:10823

OECD. (2023). *PISA 2022 Results: Factsheets. Germany, 05 December 2023,* Organisation for Economic Co-operation and Development. Verfügbar unter: https://www.oecd.org/publication/pisa-2022-results/country-notes/germany-1a2cf137/

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press.

Petty, R. E. & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion, *19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Petty, R. E. & Cacioppo, J. T. (1996). *Attitudes And Persuasion. Classic And Contemporary Approaches*. New York: Westview Press.

Pratkanis, A. R. & Greenwald, A. G. (1989). A Sociocognitive Model of Attitude Structure and Function. In L. Berkowitz (Hrsg.), *Advances in experimental social psychology* (Advances in Experimental Social Psychology, Bd. 22, S. 245–285). Orlando: Academic P. https://doi.org/10.1016/S0065-2601(08)60310-X

Putnick, D. L. & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2018). R: A Language and Environment for Statistical Computing. (Version 3.5.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Verfügbar unter: https://www.R-project.org/

Ramsteck, C. & Maier, U. (2015). Testdatenbasierte Schul- und Unterrichtsentwicklung. Analyse von Handlungsmustern bei der Rezeption und Nutzung von Vergleichsarbeitsdaten. In J. Schrader (Hrsg.), *Governance von Bildung im Wandel. Interdisziplinäre Zugänge* (Educational Governance, Bd. 28, S. 119–144). Wiesbaden: Springer VS.

Ramsteck, C., Muslic, B., Graf, T., Maier, U. & Kuper, H. (2015). Data-Based School Improvement. The Role of Principals and School Supervisory Authorities Within the Context of Low-Stakes Mandatory Proficiency Testing in Four German States. *International Journal of Educational Management*, *29*(6), 766–789. https://doi.org/10.1108/IJEM-08-2014-0109

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften*. De Gruyter Oldenbourg. https://doi.org/10.1524/9783486854008

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research (Version 1.8.12) [Computer software]. Evanston, Illinois: Northwestern University. Verfügbar unter: https://CRAN.R-project.org/package=psych

Richter, D. (2016). Die Vergleichsarbeiten in Deutschland. Eine Bestandsaufnahme. In Bundesministerium für Bildung und Forschung (Hrsg.), *Bildungsforschung 2020. Zwischen wissenschaftlicher Exzellenz und gesellschaftlicher Verantwortung. Tagung des Bundesministeriums für Bildung und Forschung vom 27. bis 28. März 2014 in Berlin* (Bildungsforschung, Bd. 42, S. 87–96).

Richter, D. & Böhme, K. (2014). Vergleichsarbeiten im Fokus: Welche Funktionen erfüllt der Test aus Sicht von Lehrkräften? *Schulmanagement.*, (2), 12–14.

Richter, D., Böhme, K., Becker, M., Pant, H. A. & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten. Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Zeitschrift für Pädagogik*, *60*(2), 225–244. https://doi.org/10.25656/01:12846

Rieß, C. & Zuber, J. (2014). *Rezeption und Nutzung von Ergebnissen der Bildungsstandardüberprüfung in Mathematik auf der 8. Schulstufe unter Berücksichtigung der Rückmeldemoderation* (2). Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (bifie).

Roeser, R. W., Mashburn, A. J., Skinner, E. A., Choles, J. R., Taylor, C., Rickert, N. P. et al. (2022). Mindfulness Training Improves Middle School Teachers' Occupational Health, Well-Being, and Interactions With Students in Their Most Stressful Classrooms. *Journal of Educational Psychology*, *114*(2), 408–425. https://doi.org/10.1037/edu0000675

Rosenberg, M. J. & Hovland, C. I. (1969). Cognitive, Affective, and Behavioral Components of Attitudes. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson & J. W. Brehm (Hrsg.), *Attitude organization and change. An analysis of consistency among attitude components.* (4. Aufl., S. 1–14). New Haven: Yale University Press.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubio, D. M. & Gillespie, D. F. (1995). Problems With Error in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(4), 367–378. https://doi.org/10.1080/10705519509540020

Sälzer, C. (2016). *Studienbuch Schulleistungsstudien. Das Rasch-Modell in der Praxis* (Mathematik im Fokus, 1. Aufl. 2016). Berlin, Heidelberg: Springer Spektrum. https://doi.org/10.1007/978-3-662-45765-8

Savalei, V. (2018). On the Computation of the RMSEA and CFI from the Mean-And-Variance Corrected Test Statistic with Nonnormal Data in SEM. *Multivariate behavioral research*, *53*(3), 419–429. https://doi.org/10.1080/00273171.2018.1455142

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989X.7.2.147

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schermelleh-Engel, K. & Schweizer, K. (2012). Multitrait-Multimethod-Analysen. In H. Moosbrugger (Hrsg.), *Testtheorie und Fragebogenkonstruktion. Mit 66 Abbildung und 41 Tabellen* (Springer-Lehrbuch, 2., aktual. und überarb. Aufl., S. 345–362). Berlin: Springer.

Schneewind, J. & Kuper, H. (2009). Rückmeldeformate und Verwendungsmöglichkeiten der Ergebnisse aus zentralen Lernstandserhebungen. In T. Bohl (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (1. Aufl., S. 113–129). Bad Heilbrunn: Klinkhardt.

Schreiber, J. B. (2008). Core Reporting Practices in Structural Equation Modeling. *Research in Social & Administrative Pharmacy: RSAP*, *4*(2), 83–97. https://doi.org/10.1016/j.sapharm.2007.04.003

Schwab, S. & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik*, *7*(3), 175–193. https://doi.org/10.25656/01:11380

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611.

Shapiro, S. S., Wilk, M. B. & Chen, H. J. (1968). A Comparative Study of Various Tests for Normality. *Journal of the American Statistical Association*, *63*(324), 1343–1372. https://doi.org/10.1080/01621459.1968.10480932

Skejic, M., Neumann, D. & Mangal, H. (2015). Vergleichsarbeiten im Fach Englisch – Einschätzungen von hessischen Lehrkräften. *Zeitschrift für Fremdsprachenforschung*, *26*(2), 183–208.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century.

Söhnchen, F. (2009). Common Method Variance und Single Source Bias. In S. Albers, D. Klapper, U. Konradt, J. Wolf & A. Walter (Hrsg.), *Methodik der empirischen Forschung* (3., überarbeitete und erweiterte Auflage, S. 137–152). Wiesbaden: Gabler Verlag.

Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik*, *35*(1), 79–99. https://doi.org/10.1007/s13138-013-0056-z

Staats, A. W. & Staats, C. K. (1958). Attitudes Established By Classical Conditioning. *The Journal of Abnormal and Social Psychology*, *57*(1), 37–40. https://doi.org/10.1037/h0042782

Stanat, P., Pant, H. A., Pöhlmann, C. & Kuhl, P. (2013). Was kann das IQB leisten? In S. Lin-Klitzing, D. DiFuccia & G. Müller-Frerich (Hrsg.), *Zur Vermessung von Schule. Empirische Bildungsforschung und Schulpraxis* (Gymnasium - Bildung - Gesellschaft, Bd. 4, S. 125–152). Bad Heilbrunn: Klinkhardt.

Steinmetz, H. (2015). *Lineare Strukturgleichungsmodelle. Eine Einführung mit R* (Sozialwissenschaftliche Forschungsmethoden, 2. Auflage). Mering: Rainer Hampp Verlag.

Steyer, R. & Eid, M. (1993). *Messen und Testen* (Springer-Lehrbuch). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-97455-7

Tabachnick, B. G. & Fidell, L. S. (2010). *Using Multivariate Statistics* (Pearson international edition, 5. Aufl.). Boston: Pearson and Allyn & Bacon.

Tarkian, J., Maritzen, N., Eckert, M. & Thiel, F. (2019). Vergleichsarbeiten (VERA) – Konzeption und Implementation in den 16 Ländern. In F. Thiel, J. Tarkian, E.-M. Lankes, N. Maritzen, T. Riecke-Baulecke & A. Kroupa (Hrsg.), *Datenbasierte Qualitätssicherung und -entwicklung in Schulen. Eine Bestandsaufnahme in den Ländern der Bundesrepublik Deutschland* (41-103). Wiesbaden, Germany: Springer VS. https://doi.org/10.1007/978-3-658-23240-5_4

Thiel, C., Schweizer, S. & Bellmann, J. (2017). Rethinking Side Effects of Accountability in Education: Insights from a Multiple Methods Study in Four German School Systems. *Education Policy Analysis Archives*, *25*(93), 1–32. https://doi.org/10.14507/epaa.25.2662

Thiel, F., Tarkian, J., Lankes, E.-M., Maritzen, N. & Riecke-Baulecke, T. (2019). Strategien datenbasierter Steuerung zur Sicherung und Entwicklung von Schulqualität in den 16 Ländern – Zusammenfassung und Diskussion. In F. Thiel, J. Tarkian, E.-M. Lankes, N. Maritzen, T. Riecke-Baulecke & A. Kroupa (Hrsg.), *Datenbasierte Qualitätssicherung und -entwicklung in Schulen. Eine Bestandsaufnahme in den Ländern der Bundesrepublik Deutschland* (S. 313–325). Wiesbaden, Germany: Springer VS. https://doi.org/10.1007/978-3-658-23240-5_8

Thurstone, L. L. (1928). Attitudes Can Be Measured. *American Journal of Sociology*, *33*(4), 529–554. https://doi.org/10.1086/214483

Thurstone, L. L. (1931). The Measurement of Social Attitudes. *The Journal of Abnormal and Social Psychology*, *26*(3), 249–269. https://doi.org/10.1037/h0070363

Tucker, L. R. & Lewis, C. (1973). A Reliability Coefficient For Maximum Likelihood Factor Analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Turner, M., Kitchenham, B., Brereton, P., Charters, S. & Budgen, D. (2010). Does the Technology Acceptance Model Predict Actual Use? A Systematic Literature Review. *Information and Software Technology*, *52*(5), 463–479. https://doi.org/10.1016/j.infsof.2009.11.005

Tuten, T. L., Urban, D. J. & Bosnjak, M. (2002). Internet Surveys and Data Quality: A Review. In B. Batinic, U.-D. Reips & M. Bosnjak (Eds.), *Online social sciences* (S. 7–27). Seattle, Wash.: Hogrefe & Huber.

Ullman, J. B. & Bentler, P. M. (2003). Structural Equation Modeling. In I. B. Weiner & J. A. Schinka (Eds.), *Handbook of Psychology. Volume 2: Research Methods in Psychology* (S. 607–634). Hoboken, NJ: Wiley.

Umbach, P. D. (2004). Web Surveys: Best Practices. *New Directions for Institutional Research*, *2004*(121), 23–38. https://doi.org/10.1002/ir.98

Urban, D. & Mayerl, J. (2014). *Strukturgleichungsmodellierung*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-01919-8

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vanhoof, J., Vanlommel, K., Thijs, S. & Vanderlocht, H. (2014). Data Use by Flemish School Principals: Impact of Attitude, Self-Efficacy and External Expectations. *Educational Studies*, *40*(1), 48–62. https://doi.org/10.1080/03055698.2013.830245

Venkatesh, V. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research*, *11*(4), 342–365. https://doi.org/10.1287/isre.11.4.342.11872

Venkatesh, V. & Davis, F. D. (1996). A Model of the Antecedents of Perceived Ease of Use: Development and Test. *Decision Sciences*, *27*(3), 451–481. https://doi.org/10.1111/j.1540-5915.1996.tb00860.x

Venkatesh, V. & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Vettorazzi, K., Emmrich, R. & Fuchs, G. (2017). Die Vergleichsarbeiten (VERA). Ein bildungsstandardbasiertes Instrument zur Unterrichts- und Schulentwicklung. In M. Brunner, R. Emmrich & H. Gärtner (Hrsg.), *ISQ-Bericht zur Schulqualität 2016. Qualitätssicherungsverfahren, Prozess- und Ergebnisqualität an Schulen in Berlin und Brandenburg* (S. 35–51). Berlin: Institut für Schulqualität der Länder Berlin und Brandenburg. Verfügbar unter: https://www.isq-bb.de/wordpress/wp-content/uploads/2017/03/ISQ_Bericht_Schulqualitaet_2016.pdf

Visscher, A. J. & Coe, R. (2002). *School Improvement Through Performance Feedback* (Contexts of learning, 1. Aufl.). London u.a.: ROUTLEDGE.

Visscher, A. J. & Coe, R. (2003). School Performance Feedback Systems: Conceptualisation, Analysis, and Reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. https://doi.org/10.1076/sesi.14.3.321.15842

Vogel, S. (2020). *Wie schätzen Lehrkräfte die Lernstandserhebungen in Mathematik ein?* Dissertation. Universität Kassel. https://doi.org/10.17170/KOBRA-202007101439

Vogel, S., Blum, W., Achmetli, K. & Krawitz, J. (2016). Qualifizierung von Lehrkräften zum konstruktiven Umgang mit zentralen Lernstandserhebungen – Ergebnisse aus dem Projekt VELM-8. *Journal für Mathematik-Didaktik*, *37*(2), 319–348. https://doi.org/10.1007/s13138-016-0092-6

Wacker, A. & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg. *Zeitschrift für Erziehungswissenschaft*, *15*(4), 683–706. https://doi.org/10.1007/s11618-012-0326-4

Wagner, I., Hosenfeld, I. & Zimmer-Müller, M. (2019). Empirische Arbeit: Vergleichende Analyse der Zusammenhänge von Akzeptanz, Auseinandersetzung mit und Nutzung von Ergebnissen von Vergleichsarbeiten und Schulinspektionen. *Psychologie in Erziehung und Unterricht*, *66*. https://doi.org/10.2378/peu2019.art22d

Wagner, I. & Koch, U. (2021). Unterschiede in der VERA-Testvorbereitung von Lehrkräften in Abhängigkeit von ihrer wahrgenommenen Funktion und ihrer Akzeptanz von Vergleichsarbeiten. *Unterrichtswissenschaft*, *49*, 373–394. https://doi.org/10.1007/s42010-021-00096-w

Warshaw, P. R. & Davis, F. D. (1985). Disentangling Behavioral Intention and Behavioral Expectation. *Journal of Experimental Social Psychology*, *21*(3), 213–228. https://doi.org/10.1016/0022-1031(85)90017-4

Weiber, R. & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS* (Springer-Lehrbuch, 2., erw. und korr. Aufl.). Berlin: Springer Gabler. https://doi.org/10.1007/978-3-642-35012-2

West, S. G., Taylor, A. B. & Wu, W. (2012). Model Fit and Model Selection in Structural Equation Modeling. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (S. 209–231). New York [u.a.]: Guilford Press.

Westfall, P. H., Henning, K. S. S. & Howell, R. D. (2012). The Effect of Error Correlation on Interfactor Correlation in Psychometric Measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 99–117. https://doi.org/10.1080/10705511.2012.634726

Wicker, A. W. (1969). Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects. *Journal of Social Issues*, *25*(4), 41–78. https://doi.org/10.1111/j.1540-4560.1969.tb00619.x

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Version 3.3.3) [Computer software]. New York: Springer-Verlag. Verfügbar unter: https://ggplot2.tidyverse.org

Wickham, H., Francois, R., Henry, L. & Müller, K. (2020). dplyr: A Grammar of Data Manipulation (Version R package version 0.8.5) [Computer software]. Verfügbar unter: https://CRAN.R-project.org/package=dplyr

Wickham, H. & Henry, L. (2019). tidyr: Tidy Messy Data. (Version R package version 1.0.0.) [Computer software]. Verfügbar unter: https://CRAN.R-project.org/package=tidyr

Wurster, S., Bach, A., Schliesing, A., Thillmann, K., Pant, H. A. & Thiel, F. (2016). Schulen als Steuerungsakteure im Bildungssystem – datenbasierte Schul- und Unterrichtsentwicklung aus der Perspektive von Schulleitungen, Fachkonferenzleitungen und Lehrkräften. In Bundesministerium für Bildung und Forschung (Hrsg.), *Steuerung im Bildungssystem. Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (Bildungsforschung, Bd. 43, S. 178–207).

Wurster, S., Feldhoff, T. & Gärtner, H. (2016). Führen verschiedene Inspektionskonzepte zu unterschiedlicher Akzeptanz und Verwendung der Ergebnisse durch Schulleitungen und

Lehrkräfte? *Zeitschrift für Erziehungswissenschaft*, *19*(3), 557–575. https://doi.org/10.1007/s11618-016-0693-3

Wurster, S. & Richter, D. (2016). Nutzung von Schülerleistungsdaten aus Vergleichsarbeiten und zentralen Abschlussprüfungen für Unterrichtsentwicklung in Brandenburger Fachkonferenzen. *Journal for educational research online*, *8*(3), 159–183. https://doi.org/10.25656/01:12820

Wurster, S., Richter, D. & Lenski, A. E. (2017). Datenbasierte Unterrichtsentwicklung und ihr Zusammenhang zur Schülerleistung. *Zeitschrift für Erziehungswissenschaft*, *20*(4), 628–650. https://doi.org/10.1007/s11618-017-0759-x

Xie, Y. (2021). knitr: A General-Purpose Package for Dynamic Report (Version R package version 1.3) [Computer software]. Verfügbar unter: https://yihui.org/knitr/

Yang, H. & Yoo, Y. (2004). It's All About Attitude: Revisiting the Technology Acceptance Model. *Decision Support Systems*, *38*(1), 19–31. https://doi.org/10.1016/S0167-9236(03)00062-9

Yuan, K.-H. & Bentler, P. M. (1998). Normal Theory Based Test Statistics in Structural Equation Modeling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309. https://doi.org/10.1111/j.2044-8317.1998.tb00682.x

Yuan, K.-H. & Bentler, P. M. (2000). Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Zajonc, R. B. (1968). Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848

Zajonc, R. B. & Markus, H. (1982). Affective and Cognitive Factors in Preferences. *Journal of Consumer Research*, *9*(2), 123–131. https://doi.org/10.1086/208905

Zanna, M. P. & Rempel, J. K. (1988). Attitudes: A New Look at an Old Concept. In D. Bar-Tal & A. W. Kruglanski (Hrsg.), *The social psychology of knowledge* (S. 315–334). Cambridge: Cambridge Univ. Pr.

Zhao, J., Fang, S. & Jin, P. (2018). Modeling and Quantifying User Acceptance of Personalized Business Modes Based on TAM, Trust and Attitude. *Sustainability*, *10*(2), 356. https://doi.org/10.3390/su10020356

Zijlstra, W. P., van der Ark, L. A. & Sijtsma, K. (2011). Outliers in Questionnaire Data. *Journal of Educational and Behavioral Statistics*, *36*(2), 186–212. https://doi.org/10.3102/1076998610366263

Zimmer-Müller, M. & Hosenfeld, I. (2013). Zehn Jahre Vergleichsarbeiten. Eine Zwischenbilanz aus verschiedenen Perspektiven. In M. Zimmer-Müller & I. Hosenfeld (Hrsg.), *Zehn*

Jahre Vergleichsarbeiten: eine Zwischenbilanz aus verschiedenen Perspektiven (Empirische Pädagogik, Bd. 27,4, S. 397–406). Landau in der Pfalz: Verlag Empirische Pädagogik.

Zimmer-Müller, M., Hosenfeld, I. & Koch, U. (2014). Rückmeldungen nach Vergleichsarbeiten in Grund- und Sekundarschulen. In H. Ditton & A. Müller (eds.), *Feedback und Rückmeldungen. Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 195–212). Münster: Waxmann.

Zinnbauer, M. & Eberl, M. (2004). Die Überprüfung von Spezifikation und Güte von Strukturgleichungsmodellen: Verfahren und Anwendung. *Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung*, (21).

Zlatkin-Troitschanskaia, O., Förster, M., Preuße, D. & Mater, O. (2016). The Relationship Between Teachers' Evidence-Based Actions and Communication, Cooperation, and Participation Structures at Schools. *Journal for educational research online*, *8*(3), 59–79. https://doi.org/10.25656/01:12806

Zuber, J. (2019). Einstellungsbildung als Gelingensbedingung für die Umsetzung einer Bildungsstandardpolitik? In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (Educational Governance, Bd. 42, S. 105–127). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-22241-3_5

Zuber, J. & Altrichter, H. (2018). The role of teacher characteristics in an educational standards reform. *Educational Assessment, Evaluation and Accountability*, *30*(2), 183–205. https://doi.org/10.1007/s11092-018-9275-7

Zumbo, B. D. (2005). Structural Equation Modeling and Test Validation. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (S. 1951–1958). Chichester, UK: John Wiley & Sons Ltd.; Wiley.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

Anhang A Ergänzungen zu Kapitel 4.2.4 Datensatzaufbereitung

Kennwerte zur Datensatzaufbereitung VERA3 2018

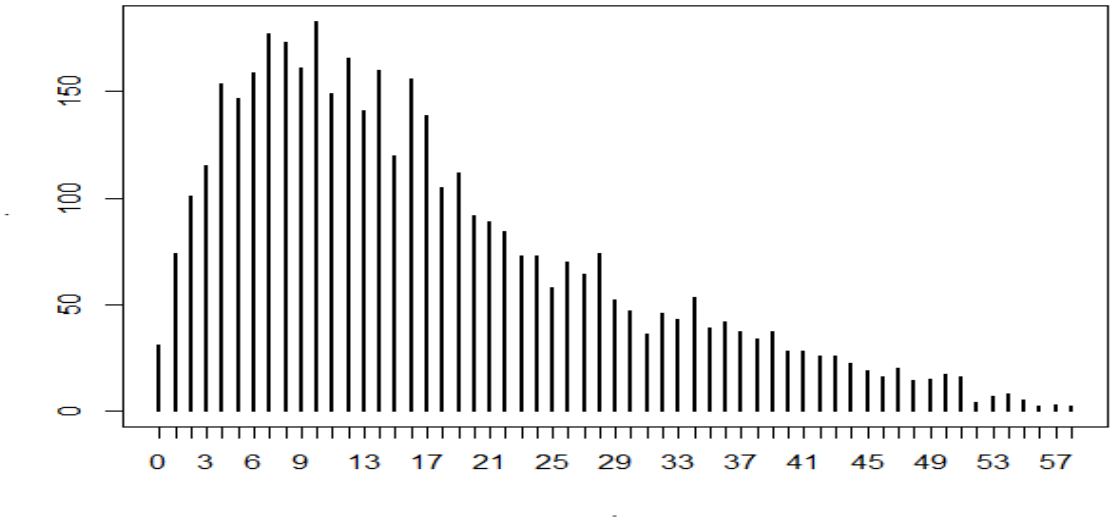


Abbildung 26: Verteilung extremer Antwortkategorien (VERA3 2018)

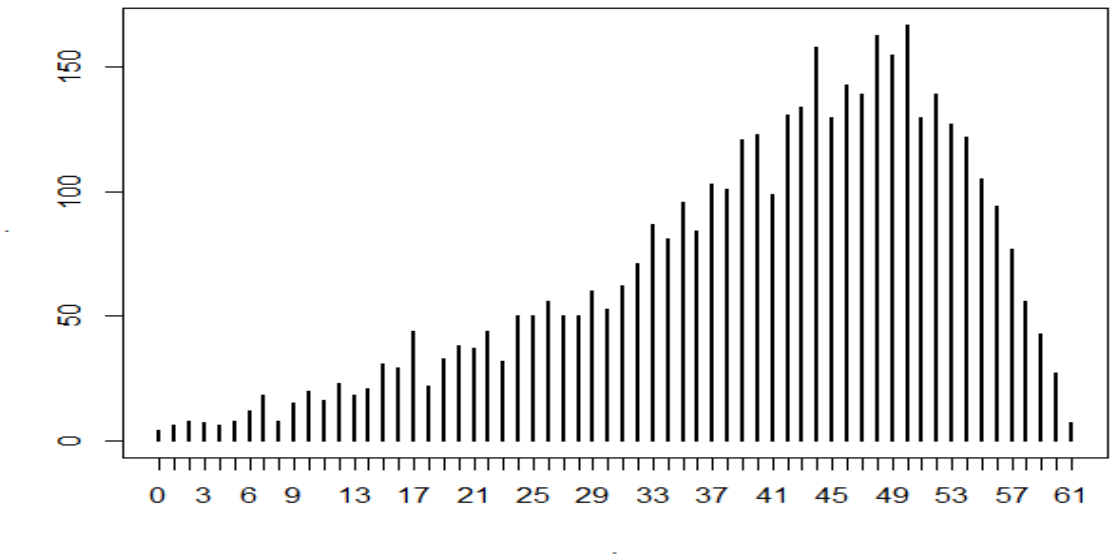


Abbildung 27: Verteilung mittlerer Antwortkategorien (VERA3 2018)

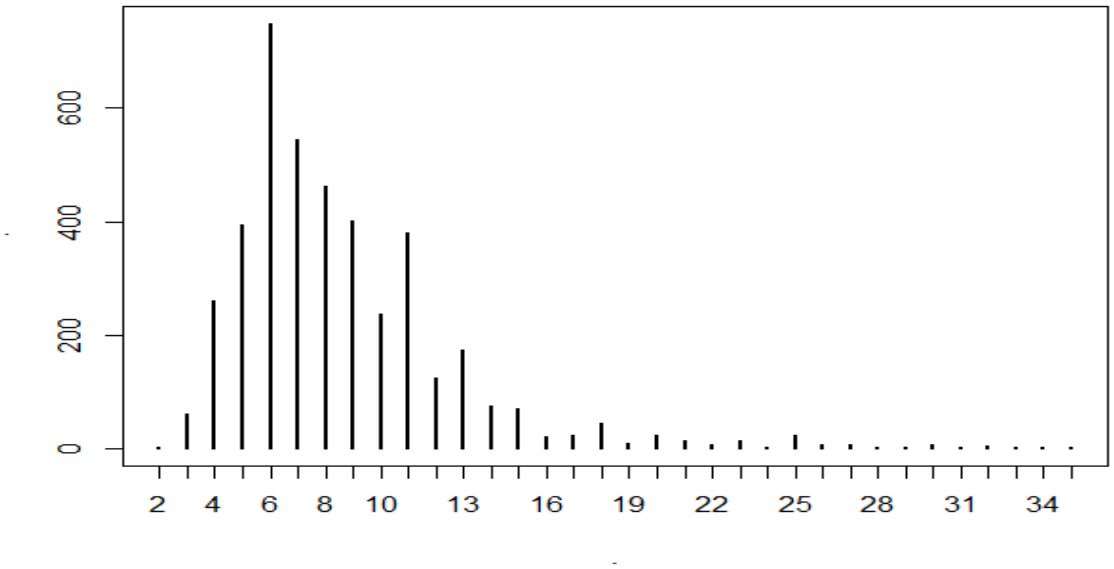


Abbildung 28: Verteilung Maximum Longstring (VERA3 2018)

Kennwerte zur Datensatzaufbereitung VERA3 2019

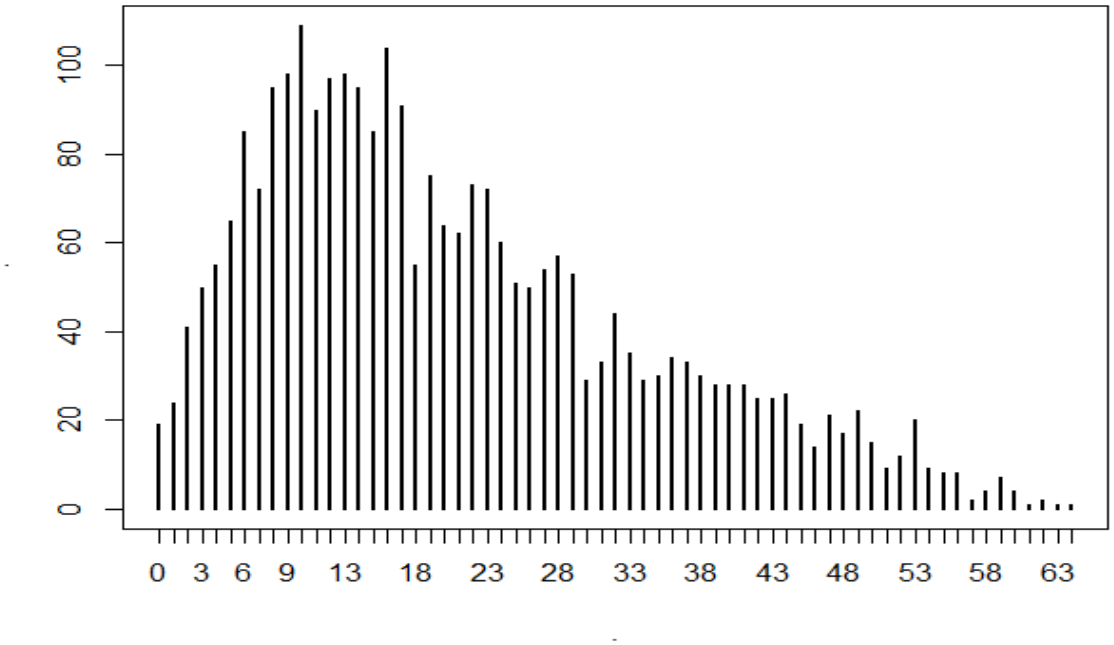


Abbildung 29: Verteilung extremer Antwortkategorien (VERA3 2019)

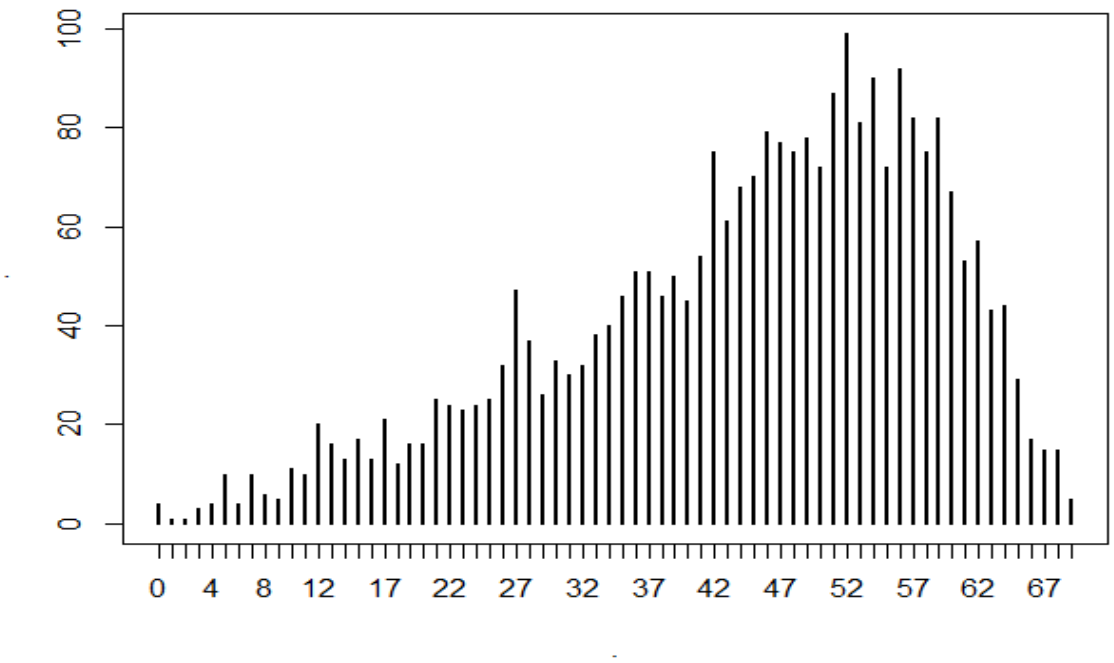


Abbildung 30: Verteilung mittlerer Antwortkategorien (VERA3 2019)

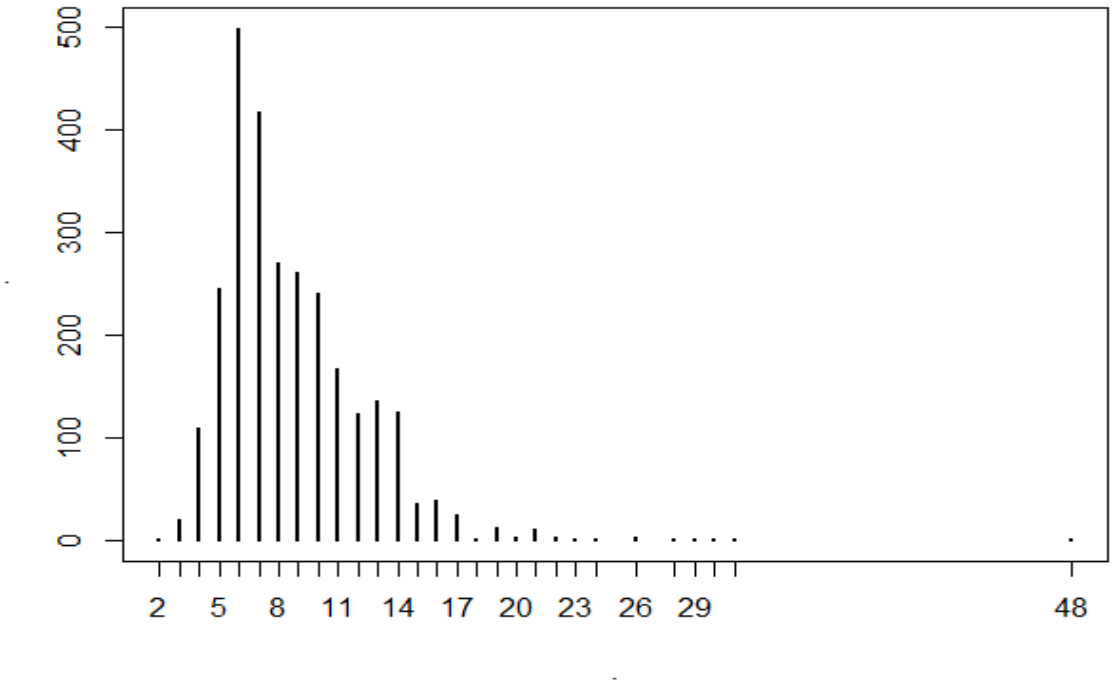


Abbildung 31: Verteilung Maximum Longstring (VERA3 2019)

Kennwerte zur Datensatzaufbereitung VERA8 2018

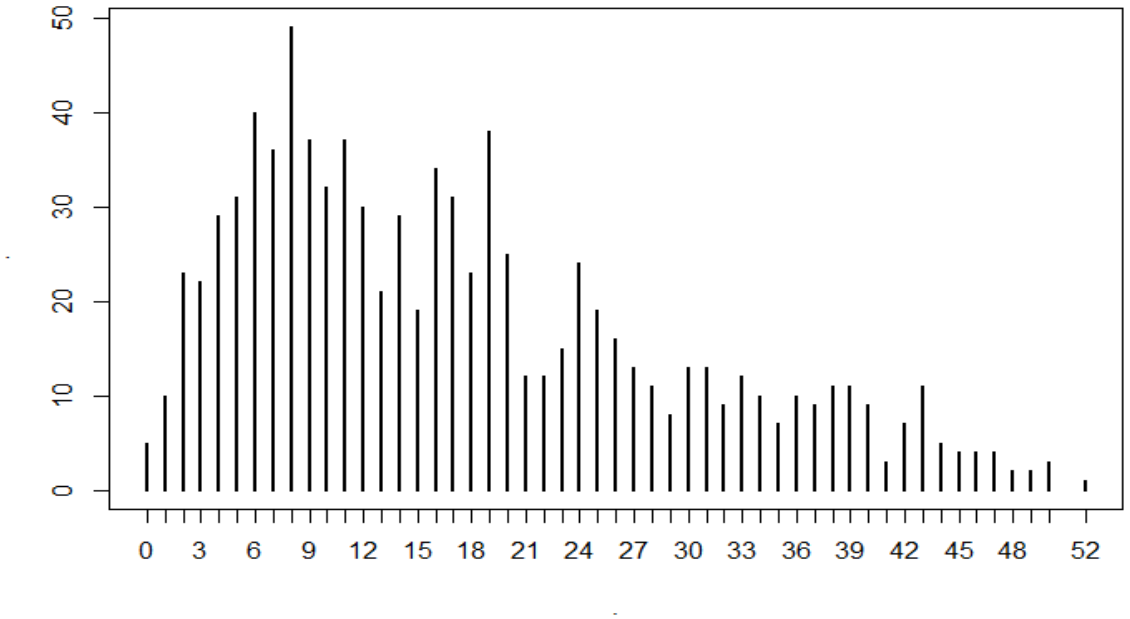


Abbildung 32: Verteilung extremer Antwortkategorien (VERA8 2018)

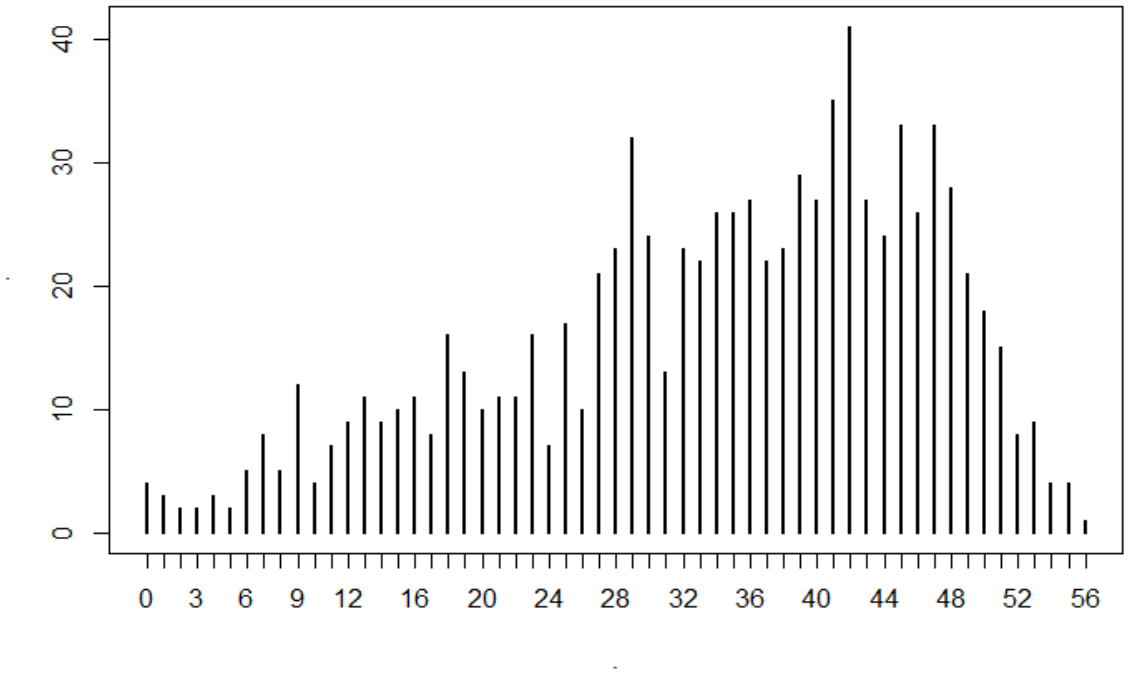


Abbildung 33: Verteilung mittlerer Antwortkategorien (VERA8 2018)

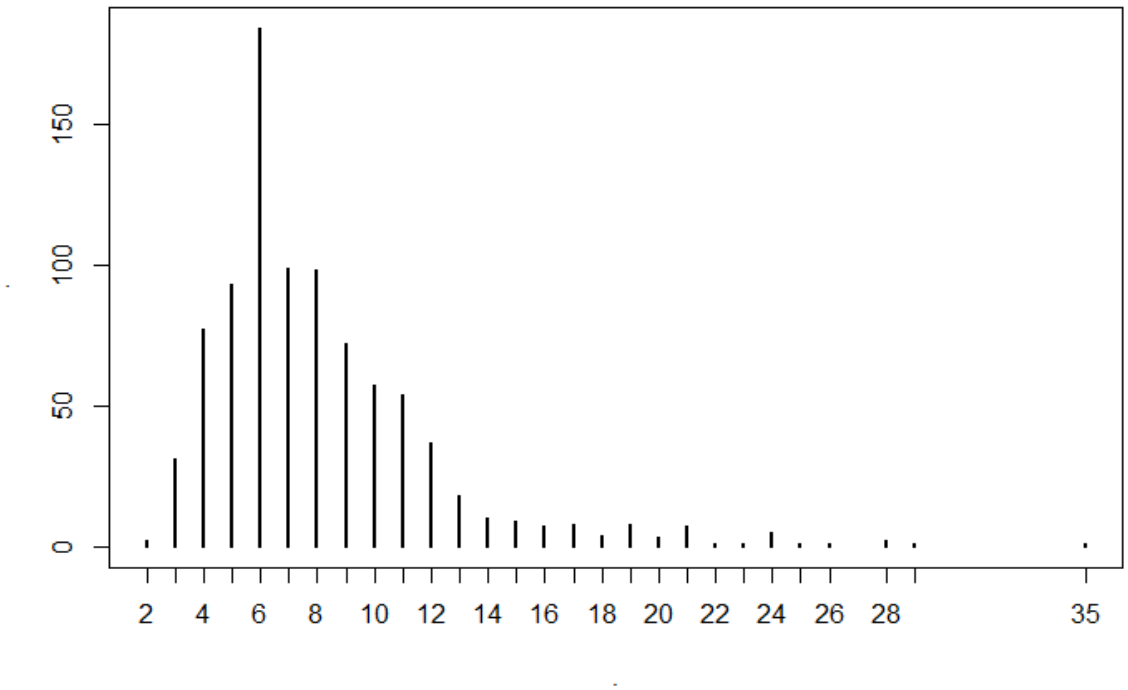


Abbildung 34: Verteilung Maximum Longstring (VERA8 2018)

Anhang B Ergänzende Ergebnisse zu Kapitel 5.1.2 Skalenanalysen – Konfirmatorische Faktorenanalyse (VERA3 2018) (CFA mit allen Konstrukten)

Tabelle 51: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen aller Faktoren basierend auf der Schätzung des gemeinsamen Messmodells

Konstrukt	Item	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
Nutzungsintention [a]	NI1	.880	.000		.774
	NI2	.839	.012	80.695	.705
	NI3	.828	.012	76.182	.686
	NI4	.707	.016	53.823	.500
	NI5	.660	.018	43.564	.436
Einstellung [a]	AE1	.929	.000		.863
	AE2	.846	.010	85.104	.716
	AE3	.903	.008	120.391	.816
	AE4	.881	.009	105.972	.777
	AE5	.841	.011	87.882	.708
Aufwand-Nutzen [a]	AN1r	.749	.000		.560
	AN2	.814	.022	47.530	.663
	AN3	.880	.019	59.437	.774
	AN4	.666	.023	43.118	.444
	AN5	.747	.022	49.863	.558
	AN6	.796	.020	61.952	.633
Nützlichkeit [a]	WN1	.807	.000		.652
	WN2	.769	.019	54.189	.591
	WN3	.835	.018	63.229	.697
	WN4	.733	.019	51.109	.538
	WN5	.783	.019	53.976	.614
Zeitliche Belastung [b]	ZB1	.403	.000		.163
	ZB2	.692	.065	20.051	.478
	ZB3	.611	.091	14.277	.373

Anmerkungen. $N = 4\,141$ (VERA3 2018); [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 52: Spezifizierte Residualkorrelationen (Modell 1$_{CFA(V3\text{-}18)_b}$)

	AN2	AN4	AN5	AN6
ZB1	-	-.39	-	-
ZB2	-	-	-.22	-
ZB3	-	-	-	-.41
AN3	.25	-	-	-
AN4	-	-	.50	.29
AN5	-	-	-	.41

Anmerkungen. Alle Koeffizienten erweisen sich als signifikant ($p < .001$).

Anhang C Ergänzende Ergebnisse zu Kapitel 5.1.3 Untersuchung der Kausalbeziehungen (VERA3 2018)

Tabelle 53: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen n der latenten Variablen in Modell $1_{(V3\text{-}18)}$ und Modell $1_{dP(V3\text{-}18)}$

		Modell $1_{(V3\text{-}18)}$				Modell $1_{dP(V3\text{-}18)}$			
Konstrukt	Item	λ^s_{ij}	*S. E.*	*z*-Wert	R^2	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
Nutzungs-intention [a]	NI1	.877	.000		.769	.880	.000		.774
	NI2	.843	.012	79.943	.710	.839	.012	80.696	.705
	NI3	.829	.012	75.110	.687	.828	.012	76.194	.686
	NI4	.707	.016	53.143	.499	.707	.016	53.838	.500
	NI5	.661	.018	43.294	.437	.661	.018	43.580	.436
Einstellung [a]	AE1	.927	.000		.859	.929	.000		.863
	AE2	.843	.010	85.491	.711	.846	.010	85.068	.716
	AE3	.900	.008	121.029	.810	.904	.008	120.351	.816
	AE4	.883	.009	107.515	.779	.881	.009	105.964	.777
	AE5	.842	.011	88.403	.708	.841	.011	87.913	.708
Aufwand-Nutzen [a]	AN1r	.749	.000		.560	.749	.000		.560
	AN2	.814	.022	47.642	.663	.814	.022	47.567	.663
	AN3	.880	.019	59.608	.774	.880	.019	59.521	.774
	AN4	.666	.023	43.083	.444	.666	.023	43.074	.444
	AN5	.747	.022	49.870	.558	.747	.022	49.840	.558
	AN6	.796	.020	62.008	.633	.796	.020	61.985	.633
Nützlichkeit [a]	WN1	.803	.000		.645	.807	.000		.652
	WN2	.775	.019	53.839	.601	.769	.019	54.231	.591
	WN3	.830	.018	62.215	.689	.835	.018	63.234	.697
	WN4	.729	.019	50.583	.531	.733	.019	51.105	.538
	WN5	.791	.020	53.496	.625	.784	.019	54.022	.614
Zeitliche Belastung [b]	ZB1	.400	.000		.160	.401	.000		.161
	ZB2	.692	.065	20.136	.478	.691	.065	20.123	.478
	ZB3	.613	.093	14.144	.375	.613	.093	14.187	.375

Anmerkungen. $N = 4\,141$ (VERA3 2018); [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 54: Spezifizierte Residualkorrelationen Modell $1_{(V3\text{-}18)}$ und Modell $1_{dP(V3\text{-}18)}$ (VERA3 2018)

	Modell $1_{(V3\text{-}18)}$				Modell $1_{dP(V3\text{-}18)}$			
	AN2	AN4	AN5	AN6	AN2	AN4	AN5	AN6
ZB1	-	-.38	-	-	-	-.39	-	-
ZB2	-	-	-.23	-	-	-	-.23	-
ZB3	-	-	-	-.41	-	-	-	-.41
AN3	.25	-	-	-	.25	-	-	-
AN4	-	-	.50	.29	-	-	.50	.29
AN5	-	-	-	.41	-	-	-	.41

Anmerkungen. Alle Koeffizienten erweisen sich als signifikant ($p < .001$).

Anhang D Ergänzende Ergebnisse zu Kapitel 5.2 Modellanpassung: Modell 2

Tabelle 55: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen der latenten Variablen in Modell 2$_{(V3-18)}$ und Modell 2$_{dP(V3-18)}$

		Modell 2$_{(V3-18)}$				Modell 2$_{dP(V3-18)}$			
Konstrukt	Item	λ^s_{ij}	*S. E.*	*z*-Wert	R^2	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
Nutzungs-intention [a]	NI1	.877	.000		.769	.879	.000		.774
	NI2	.843	.012	79.926	.710	.839	.012	80.646	.705
	NI3	.829	.012	75.114	.687	.828	.012	76.175	.686
	NI4	.706	.016	53.154	.499	.707	.016	53.855	.499
	NI5	.661	.018	43.301	.437	.661	.018	43.587	.437
Ein-stellung [a]	AE1	.927	.000		.859	.929	.000		.863
	AE2	.843	.010	84.490	.711	.846	.010	83.903	.716
	AE3	.905	.008	120.892	.819	.908	.008	119.705	.825
	AE4	.883	.009	106.350	.779	.881	.009	104.679	.777
	AE5	.835	.011	87.657	.697	.835	.011	87.247	.697
Nützlich-keit [a]	WN1	.801	.000		.642	.806	.000		.649
	WN2	.775	.019	53.445	.600	.768	.019	53.854	.590
	WN3	.833	.018	61.441	.693	.838	.018	62.547	.702
	WN4	.730	.020	50.089	.533	.735	.019	50.629	.540
	WN5	.790	.020	53.282	.624	.782	.020	53.772	.612
Zeitliche Belastung [b]	ZB1	.403	.000		.162	.403	.000		.162
	ZB2	.761	.080	18.003	.580	.760	.080	18.002	.577
	ZB3	.569	.092	13.369	.324	.570	.092	13.349	.325

Anmerkungen. N = 4 141 (VERA3 2018); [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Anhang E Ergänzende Ergebnisse zu Kapitel 5.3 Modellvalidierung (VERA3 2019)

Tabelle 56: Deskriptive Statistiken der Indikatorvariablen der latenten Konstrukte (VERA3 2019)

	Item	*M*	*SD*	Schiefe	Kurtosis	S-W-Test	Anteil fehlender Werte
Nutzungsintention Min = 1, Max = 4 M_T = 2.50	NI1	2.45	0.82	-0.17	-0.59	0.86	4.1 %
	NI2	2.49	0.83	-0.17	-0.57	0.86	4.1 %
	NI3	2.24	0.80	0.18	-0.46	0.86	4.2 %
	NI4	2.44	0.87	-0.03	-0.7	0.87	2.9 %
	NI5	2.71	0.87	-0.43	-0.44	0.85	2.5 %
Einstellung Min = 1, Max = 4 M_T = 2.50	AE1	2.39	0.95	-0.02	-0.96	0.87	1.3 %
	AE2	2.41	0.87	-0.10	-0.75	0.87	2.3 %
	AE3	2.40	0.92	-0.05	-0.90	0.87	2.1 %
	AE4	2.38	0.95	0.00	-0.98	0.87	1.1 %
	AE5	2.31	0.99	0.10	-1.09	0.87	1.4 %
Zeitliche Belastung Min = 1, Max = 5 (inv. Polung) M_T = 3.00	ZB1	2.81	0.97	-0.02	-0.09	0.89	1.6 %
	ZB2	3.05	0.80	0.07	1.21	0.82	1.6 %
	ZB3	3.43	0.85	0.28	-0.01	0.84	1.0 %
Wahrgenommene Nützlichkeit Min = 1, Max = 4 M_T = 2.50	WN1	2.30	0.81	0.08	-0.57	0.86	4.4 %
	WN2	2.27	0.89	0.15	-0.79	0.87	3.9 %
	WN3	2.35	0.88	0.00	-0.80	0.87	3.2 %
	WN4	2.50	0.91	-0.22	-0.81	0.87	3.3 %
	WN5	2.53	0.90	-0.15	-0.76	0.87	3.3 %

Anmerkungen. N = 2 751 (VERA3 2019); M_T: Theoretischer Mittelwert; inv. Polung: inverse Polung; p-Werte der Shapiro-Wilk-Tests (S-W-Test): $p < .001$.

Skalenanalyse des Konstrukts Nutzungsintention (Datengrundlage VERA3 2019)

Tabelle 57: Manifeste Korrelationen, Trennschärfe interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Nutzungsintention (VERA3 2019)

		1.	2.	3.	4.	5.	α	DEV
1.	NI1	**.86**						
2.	NI2	.76	**.84**					
3.	NI3	.73	.71	**.84**			.90	.63
4.	NI4	.64	.60	.63	**.73**			
5.	NI5	.58	.59	.58	.52	**.69**		

Anmerkungen. N = 2 751 (VERA3 2019); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 58: Fitstatistiken der konfirmatorischen Faktorenanalyse des Konstrukts Nutzungsintention (Modell $NI_{(V3\text{-}19)}$)

Modell	χ_r^2 (p-Wert)	df	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
$NI_{(V3\text{-}19)}$	20.45 (.001)	5	4.09	**.997**	**.994**	**.040**	[.023, .058]	**.009**

Anmerkungen. N = 2 751 (VERA3 2019); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Tabelle 59: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen des Faktors Nutzungsintention basierend auf der konfirmatorischen Faktorenanalyse (Modell $NI_{(V3\text{-}19)}$)

	λ^s_{ij}	*S. E.*	z-Wert	R^2
NI1	.870	.000	65.362	.756
NI2	.853	.015	58.694	.728
NI3	.837	.016	44.757	.701
NI4	.725	.020	37.402	.526
NI5	.682	.022	65.362	.466

Anmerkungen. N = 2 751 (VERA3 2019); Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Skalenanalyse des Konstrukts Einstellung (Datengrundlage VERA3 2019)

Tabelle 60: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Einstellung (VERA3 2019)

	1.	2.	3.	4.	5.	α	DEV
1. AE1	**.92**						
2. AE2	.79	**.86**					
3. AE3	.86	.80	**.91**			.95	.79
4. AE4	.81	.74	.80	**.87**			
5. AE5	.79	.76	.78	.76	**.87**		

Anmerkungen. N = 2 751 (VERA3 2019); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 61: Fitstatistiken der konfirmatorischen Faktorenanalyse des Konstrukts Einstellung (Modell $AE_{(V3\text{-}19)}$)

Modell	χ_r^2 (p-Wert)	df	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
$AE_{(V3\text{-}19)}$	18.49 (.002)	5	3.70	**.998**	**.997**	**.039**	[.021, .059]	**.005**

Anmerkungen. N = 2 751 (VERA3 2019); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Tabelle 62: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen des Faktors Einstellung basierend auf der konfirmatorischen Faktorenanalyse (Modell $AE_{(V3\text{-}19)}$)

	λ^s_{ij}	*S. E.*	z-Wert	R^2
AE1	.926	.000		.858
AE2	.863	.012	73.399	.745
AE3	.920	.009	103.369	.847
AE4	.874	.011	88.311	.765
AE5	.854	.012	79.187	.729

Anmerkungen. N = 2 751 (VERA3 2019); Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Skalenanalyse der Konstrukte Nützlichkeit und zeitliche Belastung (Datengrundlage VERA3 2019)

Tabelle 63: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Nützlichkeit (VERA3 2019)

		1.	2.	3.	4.	5.	α	DEV
1.	WN1	**.82**						
2.	WN2	.60	**.75**					
3.	WN3	.67	.67	**.83**			.88	.60
4.	WN4	.57	.50	.59	**.69**			
5.	WN5	.61	.59	.64	.54	**.77**		

Anmerkungen. N = 2 751 (VERA3 2019); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 64: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Zeitliche Belastung (VERA3 2019)

		1.	2.	3.	α	DEV
1.	ZB1	**.52**				
2.	ZB2	.36	**.66**		**.59**	.35
3.	ZB3	.22	.43	**.60**		

Anmerkungen. N = 2 751 (VERA3 2019); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 65: Fitstatistiken der konfirmatorischen Faktorenanalyse der Konstrukte Zeitliche Belastung und Nützlichkeit (Modell $WN/ZB_{(V3\text{-}19)}$)

Modell	χ_r^2 (p-Wert)	df	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR
$WN/ZB_{(V3\text{-}19)}$	79.23 (.000)	19	4.17	**.991**	**.987**	**.036**	[.028, .045]	**.031**

Anmerkungen. N = 2751 (VERA3 2019); gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben.

Tabelle 66: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen der Faktoren Nützlichkeit und Zeitliche Belastung basierend auf der konfirmatorischen Faktorenanalyse (Modell WN/ZB$_{(V3\text{-}19)}$)

	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
WN1 [a]	.794	.000		.631
WN2 [a]	.766	.024	44.570	.587
WN3 [a]	.843	.023	49.213	.711
WN4 [a]	.693	.026	38.056	.480
WN5 [a]	.767	.024	44.088	.588
ZB1 [b]	.431	.000		.185
ZB2 [b]	.821	.118	13.185	.674
ZB3 [b]	.535	.086	12.598	.286

Anmerkungen. N = 2 751 (VERA3 2019); [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 67: Korrelation der latenten Faktoren ZB und WN (Modell WN/ZB$_{(V3\text{-}19)}$)

	WN
ZB	-.23

Anmerkungen. Alle Koeffizienten erweisen sich als signifikant ($p < .001$).

Parameterschätzungen zu Modell $2_{CFA(V3-19)}$ - CFA mit allen Konstrukten (Datengrundlage VERA3 2019)

Tabelle 68: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen aller Faktoren basierend auf der Schätzung des gemeinsamen Messmodells (Modell $2_{CFA(V3-19)}$)

Konstrukt	Item	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
	NI1	.874	.000		.764
	NI2	.845	.014	67.618	.715
Nutzungsintention [a]	NI3	.833	.015	61.299	.694
	NI4	.729	.019	46.319	.532
	NI5	.685	.022	38.438	.470
	AE1	.922	.000		.850
	AE2	.863	.012	74.968	.744
Einstellung [a]	AE3	.916	.009	105.833	.839
	AE4	.879	.011	90.019	.772
	AE5	.861	.012	80.516	.741
	WN1	.838	.000		.703
	WN2	.736	.021	46.105	.541
Nützlichkeit [a]	WN3	.816	.020	51.923	.667
	WN4	.696	.024	39.448	.485
	WN5	.763	.022	46.716	.582
	ZB1	.441	.000		.194
Zeitliche Belastung [b]	ZB2	.784	.098	14.897	.615
	ZB3	.560	.091	12.252	.313

Anmerkungen. N = 2 751 (VERA3 2019). [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Strukturmodell $2_{(V3\text{-}19)}$ und $2_{dP(V3\text{-}19)}$ (Datengrundlage VERA3 2019)

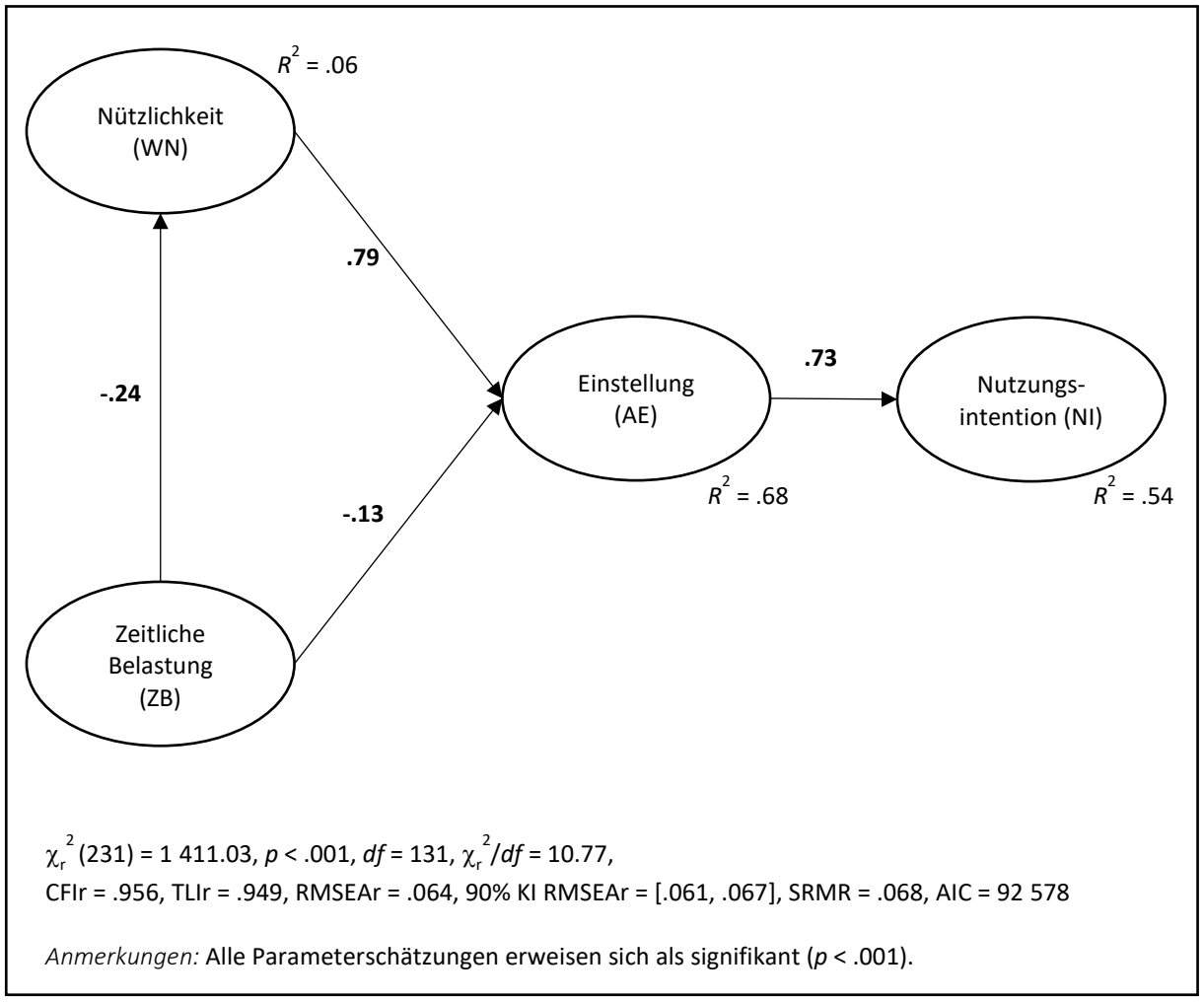
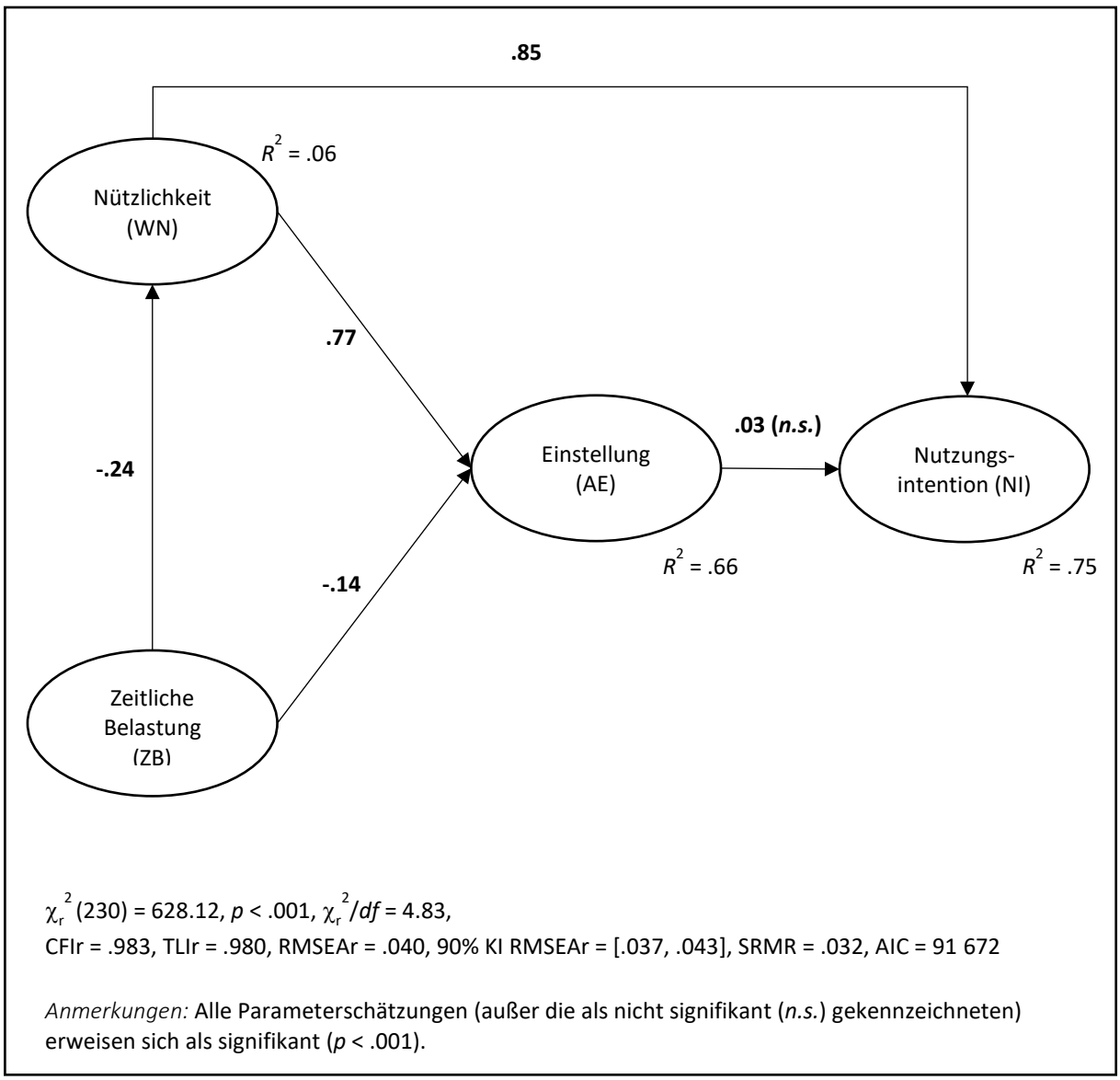


Abbildung 35: Modell $2_{(V3\text{-}19)}$ – Fitstatistiken und Parameterschätzungen (VERA3 2019) (vollständig standardisierte Lösung).



χ_r^2 (230) = 628.12, p < .001, χ_r^2/df = 4.83,
CFIr = .983, TLIr = .980, RMSEAr = .040, 90% KI RMSEAr = [.037, .043], SRMR = .032, AIC = 91 672

Anmerkungen: Alle Parameterschätzungen (außer die als nicht signifikant (*n.s.*) gekennzeichneten) erweisen sich als signifikant (p < .001).

Abbildung 36: Modell $2_{dP(V3\text{-}19)}$ – Fitstatistiken und Parameterschätzungen (VERA3 2019) (vollständig standardisierte Lösung).

Tabelle 69: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen der latenten Variablen in Modell $2_{(V3\text{-}19)}$ und Modell $2_{dP(V3\text{-}19)}$

		Modell $2_{(V3\text{-}19)}$				Modell $2_{dP(V3\text{-}19)}$			
Konstrukt	Item	λ^s_{ij}	*S. E.*	*z*-Wert	R^2	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
Nutzungs-intention [a]	NI1	.874	.000		.764	.874	.000		.764
	NI2	.846	.015	66.327	.716	.845	.014	67.633	.715
	NI3	.833	.015	60.623	.693	.833	.015	61.315	.694
	NI4	.731	.019	45.731	.534	.729	.019	46.315	.532
	NI5	.685	.022	38.109	.469	.685	.022	38.443	.470
Einstellung [a]	AE1	.918	.000		.842	.922	.000		.850
	AE2	.860	.011	75.665	.740	.863	.012	74.966	.744
	AE3	.912	.009	106.950	.831	.916	.009	105.834	.839
	AE4	.880	.011	90.744	.774	.879	.011	90.017	.772
	AE5	.863	.012	80.888	.745	.861	.012	80.516	.741
Nützlichkeit [a]	WN1	.818	.000		.670	.838	.000		.703
	WN2	.751	.022	45.692	.564	.736	.021	46.096	.541
	WN3	.822	.021	50.914	.676	.817	.020	51.924	.667
	WN4	.693	.025	38.860	.481	.696	.024	39.441	.485
	WN5	.774	.023	45.656	.599	.763	.022	46.717	.582
Zeitliche Belastung [b]	ZB1	.439	.000		.193	.440	.000		.194
	ZB2	.786	.098	14.913	.617	.785	.098	14.900	.616
	ZB3	.559	.091	12.237	.312	.559	.091	12.246	.313

Anmerkungen. $N = 2\,751$ (VERA3 2019).; [a]Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Anhang F Ergänzende Ergebnisse zu Kapitel5.4.1 Itemkennwerte und Ergebnisse der Reliabilitätsanalyse

Tabelle 70: Deskriptive Statistiken der Indikatorvariablen der latenten Konstrukte (VERA8 2018)

	Item	*M*	*SD*	Schiefe	Kurtosis	S-W-Test	Anteil fehlender Werte
Nutzungsintention Min = 1, Max = 4 M_T = 2.50	NI1	2.16	0.86	0.04	-1.03	.85	3.2 %
	NI2	2.21	0.85	-0.03	-0.98	.85	4.1 %
	NI3	2.01	0.78	0.28	-0.61	.84	4.0 %
	NI4	2.14	0.85	0.21	-0.76	.86	2.4 %
	NI5	2.49	0.91	-0.28	-0.83	.85	2.4 %
Einstellung Min = 1, Max = 4 M_T = 2.50	AE1	2.29	0.96	0.12	-1.02	.87	0.6 %
	AE2	2.39	0.85	-0.03	-0.69	.87	1.5 %
	AE3	2.37	0.93	-0.06	-0.95	.87	1.4 %
	AE4	2.25	0.93	0.12	-0.95	.87	0.9 %
	AE5	2.39	0.98	0.06	-1.03	.88	2.0 %
Zeitliche Belastung Min = 1, Max = 5 (inv. Polung) M_T = 3.00	ZB1	2.44	1.04	0.34	-0.39	.89	2.9 %
	ZB2	3.02	0.85	0.12	0.8	.84	1.5 %
	ZB3	3.88	0.88	-0.14	-0.91	.85	0.5 %
Wahrgenommene Nützlichkeit Min = 1, Max = 4 M_T = 2.50	WN1	2.18	0.89	0.19	-0.86	.86	9.1 %
	WN2	2.33	0.93	-0.04	-1.02	.86	7.5 %
	WN3	2.27	0.93	0.11	-0.94	.87	7.0 %
	WN4	2.39	0.93	-0.08	-0.94	.87	6.9 %
	WN5	2.44	0.94	-0.13	-0.95	.87	6.4 %

Anmerkungen. N = 782 (VERA8 2018); M_T: Theoretischer Mittelwert; inv. Polung: inverse Polung; p-Werte der Shapiro-Wilk-Tests (S-W-Test): $p < .001$.

Tabelle 71: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Nutzungsintention (VERA8 2018)

	1.	2.	3.	4.	5.	α	DEV
1. NI1	**.87**						
2. NI2	.76	**.86**					
3. NI3	.72	.72	**.83**			.89	.62
4. NI4	.64	.60	.59	**.72**			
5. NI5	.53	.57	.56	.49	**.67**		

Anmerkungen. $N = 782$ (VERA8 2018); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 72: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Einstellung (VERA8 2018)

	1.	2.	3.	4.	5.	α	DEV
1. AE1	**.92**						
2. AE2	.76	**.84**					
3. AE3	.84	.77	**.90**			.94	.76
4. AE4	.83	.74	.80	**.89**			
5. AE5	.76	.70	.73	.70	**.82**		

Anmerkungen. $N = 782$ (VERA8 2018); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 73: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Nützlichkeit (VERA8 2018)

		1.	2.	3.	4.	5.	α	DEV
1.	WN1	**.85**						
2.	WN2	.69	**.82**					
3.	WN3	.70	.68	**.81**			.90	.63
4.	WN4	.62	.57	.60	**.73**			
5.	WN5	.63	.66	.59	.56	**.76**		

Anmerkungen. N = 782 (VERA8 2018); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Tabelle 74: Manifeste Korrelationen, Trennschärfe, interne Konsistenz (Cronbachs Alpha) und durchschnittlich erfasste Varianz (DEV) der Skala Zeitliche Belastung (VERA8 2018)

		1.	2.	3.	α	DEV
1.	ZB1	**.53**				
2.	ZB2	.38	**.64**		.56	.31
3.	ZB3	.19	.36	**.61**		

Anmerkungen. N = 782 (VERA8 2018); alle Koeffizienten erweisen sich als signifikant ($p < .001$, zweiseitiger Test); Trennschärfe (korrigierte Item-Skala-Korrelation) in der Diagonalen.

Anhang G Ergänzende Ergebnisse zu Kapitel 5.4.2 Messinvarianzprüfung (Konstruktweise Invarianzprüfung)

Ergebnisse der Invarianzprüfung für das Konstrukt Nutzungsintention

Tabelle 75: Fitstatistiken der konfirmatorischen Faktorenanalyse für das Messmodell des Konstrukts Nutzungsintention im Gruppenvergleich zwischen VERA3 und VERA8 – Prüfung auf Messinvarianz

Modell	χ^2 (p-Wert)	df	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	ΔCFI_r	$\Delta RMSEA_r$	$\Delta SRMR$
$NI_{(V3\text{-}18,\ Teilsample)}$	43.15 (.000)	5	8.63	**.991**	**.982**	*.072*	[.053, .092]	**.019**			
$NI_{(V8\text{-}18)}$	9.38 **(.095)**	5	**1.88**	**.997**	**.994**	**.039**	[.000, .077]	**.011**			
NI_{kInv}	51.86 (.000)	10	5.19	**.993**	**.985**	*.064*	[.048, .082]	**.016**			
NI_{mInv}	60.67 (.000)	14	4.33	**.992**	**.989**	*.055*	[.041, .069]	**.020**	*-.001*	*-.009*	*.004*
NI_{skInv}	76.83 (.000)	18	4.27	**.991**	**.990**	*.053*	[.041, .066]	**.023**	*-.001*	*-.002*	*.003*
NI_{strInv}	83.07 (.000)	23	3.61	**.990**	**.991**	**.049**	[.049, .038]	**.025**	*-.001*	*-.004*	*.002*

Anmerkungen. VERA8 2018: N = 782; VERA3 2018-Teilstichprobe: N = 1 918; gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben;
Modell $NI_{(V3\text{-}18,\ Teilsample)}$: Messmodell Nutzungsintention VERA3;
Modell $NI_{(V8\text{-}18)}$: Messmodell Nutzungsintention VERA8;
Modell NI_{kInv}: konfigurale Invarianz;
Modell NI_{mInv}: metrische Invarianz;
Modell NI_{skInv}: skalare Invarianz;
Modell NI_{strInv}: strikte Invarianz.

Tabelle 76: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz für Modell NI_{strInv} (strikte Invarianz) der konfirmatorischen Faktorenanalyse der Indikatorvariablen des Konstrukts Nutzungsintention im Gruppenvergleich zwischen VERA3 und VERA8

		VERA3				VERA8			
Konstrukt	Item	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
Nutzungs-intention [a]	NI1	.882	.000		.777	.886	.000		.786
	NI2	.856	.014	67.792	.734	.862	.014	67.792	.743
	NI3	.842	.014	61.030	.709	.848	.014	61.030	.718
	NI4	.696	.019	42.835	.484	.704	.019	42.835	.496
	NI5	.642	.023	33.180	.412	.651	.023	33.180	.424

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; [a] Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 77: Fitstatistiken der konfirmatorischen Faktorenanalyse für das Messmodell des Konstrukts Einstellung im Gruppenvergleich zwischen VERA3 und VERA8 – Prüfung auf Messinvarianz

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	ΔCFI_r	$\Delta RMSEA_r$	ΔSRMR
$AE_{(V3\text{-}18,\ Teilsample)}$	21.92 (.000)	5	4.38	**.997**	**.994**	*.054*	[.032, .079]	**.007**			
$AE_{(V8\text{-}18)}$	7.44 **(.190)**	5	**1.49**	**.999**	**.998**	**.032**	[.000, .077]	**.007**			
AE_{kInv}	44.74 (.000)	10	4.47	**.997**	**.995**	**.048**	[.034, .063]	**.006**			
AE_{mInv}	57.98 (.000)	14	4.14	**.997**	**.996**	**.043**	[.032, .054]	**.011**	*.000*	*.005*	*-.005*
AE_{skInv}	111.52 (.000)	18	6.20	**.995**	**.994**	*.053*	[.044, .062]	**.015**	*-.002*	*-.010*	*-.004*
AE_{strInv}	116.30 (.000)	23	5.06	**.994**	**.995**	**.050**	[.041, .059]	**.016**	*-.001*	*.003*	*-.001*

Anmerkungen. VERA8 2018: *N* = 782; VERA3 2018-Teilstichprobe: *N* = 1 918; gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben;
Modell $AE_{(V3\text{-}18,\ Teilsample)}$: Messmodell Nutzungsintention VERA3;
Modell $AE_{(V8\text{-}18)}$: Messmodell Nutzungsintention VERA8;
Modell AE_{kInv}: konfigurale Invarianz;
Modell AE_{mInv}: metrische Invarianz;
Modell AE_{skInv}: skalare Invarianz;
Modell AE_{strInv}: strikte Invarianz.

Tabelle 78: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz für Modell AE_{strInv} (strikte Invarianz) der konfirmatorischen Faktorenanalyse der Indikatorvariablen des Konstrukts Einstellung im Gruppenvergleich zwischen VERA3 und VERA8

		VERA3				VERA8			
Konstrukt	Item	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
Einstellung [a]	AE1	.928	.000		.861	.932	.000		.869
	AE2	.847	.010	89.366	.717	.855	.010	89.366	.731
	AE3	.912	.008	125.263	.832	.917	.008	125.263	.841
	AE4	.874	.008	110.744	.763	.881	.008	110.744	.776
	AE5	.824	.010	91.635	.679	.833	.010	91.635	.694

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; [a] Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 79: Fitstatistiken der konfirmatorischen Faktorenanalyse für das Messmodell des Konstrukts Nützlichkeit im Gruppenvergleich zwischen VERA3 und VERA8 – Prüfung auf Messinvarianz

Modell	χ_r^2 (*p*-Wert)	*df*	χ_r^2/df	CFI_r	TLI_r	$RMSEA_r$	90 % KI $RMSEA_r$	SRMR	ΔCFI_r	$\Delta RMSEA_r$	ΔSRMR
$WN_{(V3\text{-}18,\ Teilsample)}$	10.58 (.000)	5	*2.12*	**.998**	**.997**	**.029**	[NA, .054]	**.008**			
$WN_{(V8\text{-}18)}$	12.88 (.000)	5	*2.58*	**.995**	**.989**	*.053*	[.018, .090]	**.013**			
WN_{kInv}	23.38 (.000)	10	*2.34*	**.997**	**.995**	**.038**	[.018, .058]	**.009**			
WN_{mInv}	39.16 (.000)	14	*2.80*	**.995**	**.993**	**.042**	[.026, .057]	**.023**	*-.002*	*.004*	*.014*
WN_{skInv}	67.78 (.000)	18	3.77	**.991**	**.990**	.051	[.038, .064]	**.026**	*-.004*	*.009*	*.003*
WN_{strInv}	78.38 (.000)	23	3.41	**.990**	**.991**	**.048**	[.036, .060]	**.024**	*-.001*	*-.003*	*-.002*

Anmerkungen. VERA8 2018: *N* = 782; VERA3 2018-Teilstichprobe: *N* = 1 918; gute Kennwerte sind fett, akzeptable Kennwerte kursiv hervorgehoben;
Modell $WN_{(V3\text{-}18,\ Teilsample)}$: Messmodell Nutzungsintention VERA3;
Modell $WN_{(V8\text{-}18)}$: Messmodell Nutzungsintention VERA8;
Modell WN_{kInv}: konfigurale Invarianz;
Modell WN_{mInv}: metrische Invarianz;
Modell WN_{skInv}: skalare Invarianz;
Modell WN_{strInv}: strikte Invarianz.

Tabelle 80: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz für Modell WN_{strInv} (strikte Invarianz) der konfirmatorischen Faktorenanalyse der Indikatorvariablen des Konstrukts Nützlichkeit im Gruppenvergleich zwischen VERA3 und VERA8

		VERA3				VERA8			
Konstrukt	Item	λ^s_{ij}	*S. E.*	*z*-Wert	R^2	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
Nützlich-keit [a]	WN1	.805	.000		.647	.817	.000		.668
	WN2	.777	.023	44.208	.604	.791	.023	44.208	.625
	WN3	.835	.022	50.194	.697	.846	.022	50.194	.716
	WN4	.721	.024	39.042	.520	.737	.024	39.042	.543
	WN5	.784	.024	43.844	.614	.797	.024	43.844	.635

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; [a] Wertebereich der Variablen jeweils 1 bis 4; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Ergänzende Ergebnisse der Invarianzprüfung für die Konstrukte Nützlichkeit und zeitliche Belastung

Tabelle 81: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz für Modell WN/ZB$_{strInv\text{-}partiell}$ (partielle strikte Invarianz) der konfirmatorischen Faktorenanalyse der Indikatorvariablen des Konstrukts Nützlichkeit und zeitliche Belastung im Gruppenvergleich zwischen VERA3 und VERA8

		VERA3				VERA8			
Konstrukt	Item	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2	λ^{s}_{ij}	*S. E.*	*z*-Wert	R^2
Nützlichkeit [a]	WN1	.804	.000		.647	.816	.000		.666
	WN2	.777	.023	44.337	.603	.790	.023	44.337	.624
	WN3	.835	.022	50.536	.697	.845	.022	50.536	.715
	WN4	.722	.024	39.159	.521	.737	.024	39.159	.543
	WN5	.785	.024	43.810	.616	.798	.024	43.810	.636
Zeitliche Belastung [b]	ZB1	.398	.000		.159	.495	.000		.245
	ZB2	.722	.113	11.908	.521	.808	.113	11.908	.653
	ZB3	.502	.092	11.214	.252	.555	.092	11.214	.308

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 82: Latente Faktorkorrelationen der Konstrukte Nützlichkeit und zeitliche Belastung Modell WN/ZB$_{strInv\text{-}partiell}$ (partielle strikte Invarianz)

	Zeitliche Belastung	
	VERA3	VERA8
Nützlichkeit	-.29	-.41

Anmerkungen. Alle Koeffizienten erweisen sich als signifikant ($p < .001$).

Anhang H Ergänzende Ergebnisse zu Kapitel 5.4.2 Messinvarianzprüfung (CFA mit allen Konstrukten)

Tabelle 83: Unstandardisierte und standardisierte Intercepts von Modell GroupCFA$_{strInv\text{-}partiell\ b}$ (partiell strikte Invarianz)

		VERA3			VERA8		
Konstrukt	Item	Unst.	*S. E.*	Std.	Unst.	*S. E.*	Std.
Nutzungs-intention [a]	NI1	2.338	.019	2.798	2.338	.019	2.745
	NI2	2.371	.018	2.892	2.371	.018	2.843
	NI3	2.148	.017	2.779	2.148	.017	2.733
	NI4	2.330	.018	2.747	2.330	.018	2.715
	NI5	2.667	.018	3.071	2.667	.018	3.040
Einstellung [a]	AE1	2.401	.020	2.658	2.401	.020	2.538
	AE2	2.473	.018	3.020	2.473	.018	2.910
	AE3	2.455	.020	2.786	2.455	.020	2.666
	AE4	2.383	.020	2.687	2.383	.020	2.576
	AE5	2.417	.020	2.594	2.417	.020	2.503
Nützlichkeit [a]	WN1	2.389	.019	2.828	2.389	.019	2.741
	WN2	2.402	.019	2.710	2.402	.019	2.637
	WN3	2.446	.020	2.701	2.446	.020	2.618
	WN4	2.516	.019	2.814	2.516	.019	2.747
	WN5	2.630	.020	2.880	2.630	.020	2.801
Zeitliche Belastung [b]	ZB1	2.636	.018	2.818	2.636	.018	2.674
	ZB2	3.036	.014	4.338	3.036	.014	3.754
	ZB3	3.397	.018	4.416	3.910	.030	4.285

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; Unst.: unstandardisiert; Std.: standardisiert; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 84: Unstandardisierte und standardisierte Messfehlervarianzen von Modell GroupCFA$_{strInv\text{-}partiell\ b}$ (partiell strikte Invarianz)

		VERA3			VERA8		
Konstrukt	Item	Unst.	*S. E.*	Std.	Unst.	*S. E.*	Std.
Nutzungs-intention [a]	NI1	.147	.007	.211	.147	.007	.203
	NI2	.191	.009	.284	.191	.009	.274
	NI3	.179	.008	.299	.179	.008	.289
	NI4	.364	.014	.506	.364	.014	.494
	NI5	.442	.013	.586	.442	.013	.574
Einstellung [a]	AE1	.111	.006	.136	.111	.006	.124
	AE2	.210	.009	.313	.210	.009	.290
	AE3	.138	.008	.177	.138	.008	.162
	AE4	.168	.008	.214	.168	.008	.197
	AE5	.292	.012	.336	.292	.012	.313
Nützlichkeit [a]	WN1	.229	.011	.321	.229	.011	.301
	WN2	.323	.012	.411	.323	.012	.390
	WN3	.260	.011	.317	.260	.011	.298
	WN4	.380	.014	.476	.380	.014	.454
	WN5	.330	.013	.396	.330	.013	.374
Zeitliche Belastung [b]	ZB1	.740	.026	.845	.740	.026	.761
	ZB2	.260	.030	.531	.260	.030	.398
	ZB3	.422	.026	.714	.542	.045	.651

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; Unst.: unstandardisiert; Std.: standardisiert; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Anhang I Ergänzende Ergebnisse zu Kapitel 5.4.4 Analyse des Strukturmodells

Parameterschätzungen zu Modell 2 Group(PKg)

Tabelle 85: Standardisierte Faktorladungen, Standardfehler, z-Werte und aufgeklärte Varianz der Indikatorvariablen in StrukturModell 2 Group$_{(PKg)}$ (partielle strikte Invarianz) im Gruppenvergleich zwischen VERA3 und VERA8

		VERA3				VERA8			
Konstrukt	Item	λ^s_{ij}	*S. E.*	*z*-Wert	R^2	λ^s_{ij}	*S. E.*	*z*-Wert	R^2
Nutzungs-intention [a]	NI1	.888	.000		.788	.885	.000		.783
	NI2	.853	.013	70.674	.727	.849	.013	70.674	.721
	NI3	.841	.014	63.228	.707	.837	.014	63.228	.701
	NI4	.707	.018	44.482	.500	.701	.018	44.482	.492
	NI5	.649	.022	34.196	.421	.643	.022	34.196	.413
Ein-stellung [a]	AE1	.928	.000		.861	.931	.000		.866
	AE2	.829	.013	64.245	.688	.835	.013	64.245	.697
	AE3	.904	.010	98.271	.818	.908	.010	98.271	.824
	AE4	.890	.010	89.621	.791	.894	.010	89.621	.799
	AE5	.818	.014	65.417	.669	.824	.014	65.417	.678
Nützlich-keit [a]	WN1	.822	.000		.675	.833	.000		.694
	WN2	.773	.022	45.784	.598	.787	.022	45.784	.619
	WN3	.819	.020	53.425	.671	.831	.020	53.425	.690
	WN4	.721	.023	41.018	.520	.736	.023	41.018	.542
	WN5	.785	.023	45.800	.616	.797	.023	45.800	.636
Zeitliche Belastung [b]	ZB1	.390	.000		.152	.490	.000		.240
	ZB2	.680	.089	14.598	.463	.776	.089	14.598	.603
	ZB3	.535	.123	9.170	.286	.595	.123	9.170	.354

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 86: Unstandardisierte und standardisierte Intercepts von StrukturModell 2 Group$_{(PKg)}$ (partiell strikte Invarianz, gleich gesetzte Pfadkoeffizienten)

		VERA3			VERA8		
Konstrukt	Item	Unst.	*S. E.*	Std.	Unst.	*S. E.*	Std.
Nutzungs-intention [a]	NI1	2.339	.019	2.769	2.339	.019	2.804
	NI2	2.372	.018	2.865	2.372	.018	2.899
	NI3	2.148	.017	2.753	2.148	.017	2.785
	NI4	2.331	.018	2.729	2.331	.018	2.751
	NI5	2.667	.018	3.054	2.667	.018	3.075
Einstellung [a]	AE1	2.402	.020	2.635	2.402	.020	2.585
	AE2	2.473	.018	2.999	2.473	.018	2.954
	AE3	2.455	.020	2.763	2.455	.020	2.714
	AE4	2.383	.020	2.666	2.383	.020	2.620
	AE5	2.417	.020	2.577	2.417	.020	2.539
Nützlichkeit [a]	WN1	2.387	.019	2.825	2.387	.019	2.742
	WN2	2.401	.019	2.709	2.401	.019	2.638
	WN3	2.444	.020	2.698	2.444	.020	2.619
	WN4	2.515	.019	2.812	2.515	.019	2.748
	WN5	2.629	.020	2.879	2.629	.020	2.801
Zeitliche Belastung [b]	ZB1	2.636	.018	2.820	2.636	.018	2.670
	ZB2	3.036	.014	4.347	3.036	.014	3.739
	ZB3	3.397	.018	4.424	3.910	.030	4.270

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; Unst.: unstandardisiert; Std.: standardisiert; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 87: Unstandardisierte und standardisierte Messfehlervarianzen von Strukturmodell 2 Group$_{(PKg)}$ (partiell strikte Invarianz, gleichgesetzte Pfadkoeffizienten)

		VERA3			VERA8		
Konstrukt	Item	Unst.	*S. E.*	Std.	Unst.	*S. E.*	Std.
Nutzungs-intention [a]	NI1	.151	.007	.212	.151	.007	.217
	NI2	.187	.009	.273	.187	.009	.279
	NI3	.178	.008	.293	.178	.008	.299
	NI4	.365	.014	.500	.365	.014	.508
	NI5	.441	.013	.579	.441	.013	.587
Einstellung [a]	AE1	.116	.006	.139	.116	.006	.134
	AE2	.212	.009	.312	.212	.009	.303
	AE3	.144	.008	.182	.144	.008	.176
	AE4	.167	.007	.209	.167	.007	.201
	AE5	.291	.012	.331	.291	.012	.322
Nützlichkeit [a]	WN1	.232	.011	.325	.232	.011	.306
	WN2	.316	.012	.402	.316	.012	.381
	WN3	.270	.011	.329	.270	.011	.310
	WN4	.383	.014	.480	.383	.014	.458
	WN5	.321	.013	.384	.321	.013	.364
Zeitliche Belastung [b]	ZB1	.741	.026	.848	.741	.026	.760
	ZB2	.262	.029	.537	.262	.029	.397
	ZB3	.421	.025	.714	.542	.043	.646

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; Unst.: unstandardisiert; Std.: standardisiert; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Parameterschätzungen zu Modell 2_{dP} Group$_{(PKg)}$

Tabelle 88: Unstandardisierte und standardisierte Intercepts von Strukturmodell 2_{dP} Group$_{(PKg)}$ (partiell strikte Invarianz, gleichgesetzte Pfadkoeffizienten)

		VERA3			VERA8		
Konstrukt	Item	Unst.	*S. E.*	Std.	Unst.	*S. E.*	Std.
Nutzungs-intention [a]	NI1	2.339	.019	2.781	2.339	.019	2.783
	NI2	2.371	.018	2.876	2.371	.018	2.878
	NI3	2.148	.017	2.764	2.148	.017	2.766
	NI4	2.331	.018	2.737	2.331	.018	2.738
	NI5	2.667	.018	3.061	2.667	.018	3.063
Einstellung [a]	AE1	2.401	.020	2.638	2.401	.020	2.580
	AE2	2.473	.018	3.002	2.473	.018	2.948
	AE3	2.455	.020	2.766	2.455	.020	2.708
	AE4	2.383	.020	2.669	2.383	.020	2.615
	AE5	2.417	.020	2.579	2.417	.020	2.535
Nützlichkeit [a]	WN1	2.389	.019	2.827	2.389	.019	2.744
	WN2	2.402	.019	2.709	2.402	.019	2.640
	WN3	2.446	.020	2.700	2.446	.020	2.621
	WN4	2.516	.019	2.814	2.516	.019	2.750
	WN5	2.630	.020	2.879	2.630	.020	2.803
Zeitliche Belastung [b]	ZB1	2.636	.018	2.819	2.636	.018	2.670
	ZB2	3.036	.014	4.347	3.036	.014	3.740
	ZB3	3.397	.018	4.424	3.910	.030	4.270

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; Unst.: unstandardisiert; Std.: standardisiert; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Tabelle 89: Unstandardisierte und standardisierte Messfehlervarianzen von Strukturmodell 2_{dP} $Group_{(PKg)}$ (partiell strikte Invarianz, gleichgesetzte Pfadkoeffizienten)

		VERA3			VERA8		
Konstrukt	Item	Unst.	*S. E.*	Std.	Unst.	*S. E.*	Std.
Nutzungs-intention [a]	NI1	.148	.007	.209	.148	.007	.209
	NI2	.191	.009	.281	.191	.009	.281
	NI3	.179	.008	.296	.179	.008	.296
	NI4	.364	.014	.502	.364	.014	.503
	NI5	.441	.013	.581	.441	.013	.582
Einstellung [a]	AE1	.111	.006	.134	.111	.006	.128
	AE2	.210	.009	.309	.210	.009	.298
	AE3	.138	.008	.175	.138	.008	.168
	AE4	.168	.008	.211	.168	.008	.203
	AE5	.292	.012	.332	.292	.012	.321
Nützlichkeit [a]	WN1	.229	.011	.320	.229	.011	.302
	WN2	.323	.012	.411	.323	.012	.390
	WN3	.261	.011	.317	.261	.011	.299
	WN4	.380	.014	.476	.380	.014	.454
	WN5	.330	.013	.395	.330	.013	.375
Zeitliche Belastung [b]	ZB1	.741	.026	.848	.741	.026	.760
	ZB2	.262	.029	.537	.262	.029	.397
	ZB3	.421	.025	.714	.542	.043	.646

Anmerkungen. VERA8 2018: $N = 782$; VERA3 2018-Teilstichprobe: $N = 1\,918$; Unst.: unstandardisiert; Std.: standardisiert; [a] Wertebereich der Variablen jeweils 1 bis 4; [b] Wertebereich der Variablen jeweils 1 bis 5; alle Parameterschätzungen erweisen sich als signifikant ($p < .001$).

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die Dissertation selbst angefertigt und alle von mir benutzten Quellen und Hilfsmittel in der Arbeit angegeben habe.

Ich habe die Dissertation bislang weder als Gesamtschrift noch in Auszügen als Prüfungsarbeit für eine staatliche oder eine andere wissenschaftliche Prüfung eingereicht.

Zudem erkläre ich, dass ich weder diese noch eine andere Abhandlung bisher bei einer anderen Hochschule als Dissertation eingereicht habe.

Mannheim, den ´ Johanna Detzel

Lebenslauf

PERSÖNLICHE DATEN

Name Johanna Detzel, geb. Siegk

BERUFLICHE TÄTIGKEITEN

Seit 01/2016 **Wissenschaftliche Mitarbeiterin – Projekt VERA**
Zentrum für Empirische Pädagogische Forschung (zepf),
Rheinland-Pfälzische Technische Universität (RPTU), Campus Landau
Dissertationsprojekt
Fachbereich: Pädagogische Psychologie
Titel: „Vergleichsarbeiten (VERA) – Eine empirische Untersuchung zur Akzeptanz bei Lehrkräften“

05/2015 – 12/2015 **Wissenschaftliche Mitarbeiterin**
Hochschule der Bundesagentur für Arbeit (HdBA)

10/2014 – 12/2014 **Selbstständige Tätigkeit – Qualitative Forschung**
Inhaltsanalytische Auswertungen transkribierter Berufsberatungsgespräche für ein Forschungsprojekt der Hochschule der Bundesagentur für Arbeit (HdBA)

04/2014 – 09/2014 **Praktikum**
Hochschule der Bundesagentur für Arbeit (HdBA)
Forschungs- und Entwicklungsprojekt BET-U25 (s.o.)

AUSBILDUNG

09/2010 - 01/2014 **Masterstudium der Wirtschaftspädagogik (M.Sc.)**
Universität Mannheim
Abschlussnote: 1,8
Empirische Masterarbeit:
„Reverse Innovation: Eine empirische Untersuchung der Wahrnehmung und der Einführung in entwickelten Ländern“ am Lehrstuhl für Marketing & Innovation (Note: 1,3)

09/2007 - 09/2010 **Bachelorstudium der Wirtschaftspädagogik (B.Sc.)**
Universität Mannheim
Abschlussnote: 2,8
Bachelorarbeit:
„Multichannel Retailing: Der Einfluss auf das Kaufverhalten“
am Lehrstuhl für Marketing & Innovation (Note: 1,3)

08/1998 - 03/2007 **Abitur**
Theodor-Heuss-Gymnasium in Ludwigshafen

KENNTNISSE & FÄHIGKEITEN

Deutsch	- Muttersprache
Englisch	- sehr gute Kenntnisse
Französisch	- Grundkenntnisse
Latein	- Latinum
MS-Office:	
-Word, Excel, PowerPoint	- sehr gute Kenntnisse
SPSS	- sehr gute Kenntnisse
R	- gute Kenntnisse
MaxQDA	- gute Kenntnisse
Limesurvey	- gute Kenntnisse

SONSTIGE QUALIFIKATIONEN

04/2011 Berufs- und arbeitspädagogische Qualifikation (AdA-Schein)

PUBLIKATIONEN UND KONFERENZEN

Siegk, J., Rübner, M., Höft, S., Sauer, S. (2015, Juli). Change of career choice readiness through career guidance: A pre-post design study. Paper Presentation, European Congress of Psychology (ECP). Mailand, 10.07.2015.

Bösinger-Schmidt, M., Höft. S., Rübner, M., Sauer, S., Siegk, S. (2015, Juli). Assessment of career choice readiness by self-report, counselors' and independent observers' ratings: A triangulation approach. Poster Presentation, European Congress of Psychology (ECP). Mailand, 10.07.2015.

Rübner, M., Höft, S., Siegk, J., Sauer, S. (2015, Juli). Beratungs- und Evaluationstool zur Erfassung der individuellen Wirkung von beruflicher Beratung für den Personenkreis U 25. ELGPN-Seminar zu „Wirkungen, Nutzen und Evidenzbasierung lebensbegleitender Beratung". Nationales Forum Beratung in Bildung, Beruf und Beschäftigung (nfb) & Bundesministerium für Bildung und Forschung. Berlin, 02.07.2015.

Schuhmacher, Monika C., Kuester, Sabine, Siegk, Johanna (2014), "Reverse Innovations: What to Communicate when Launching Innovations from Emerging Markets to Developed Markets", in Proceedings of the 21st International Product Development Management Conference, Limerick, Irland (16.-17. Juni 2014), 209-210.

Rübner, M., Höft, S., Sauer, S., Bösinger-Schmidt, M., Siegk, J. (2014): Beratungseffekte sichtbar machen. Entwicklung eines Selbstevaluationstools zur Erfassung der individuellen Wirkung von beruflicher Beratung für den Personenkreis U 25 / Berufsberatung - Sekundarstufe I (BET-U25). Newsletter Nationales Forum Beratung in Bildung, Beruf und Beschäftigung, 3, S. 2-4. Abrufbar unter: http://www.forum-beratung.de/cms/upload/Veroeffentlichungen/Newsletter/nfb-Newsletter_03-2014_fin.pdf

Mannheim, den 22.01.2024