Scharfenberg, Jonas; Keller-Schneider, Manuela; Weiß, Sabine; Hellsten, Meeri; Kiel, Ewald

Konstruktion von Vergleichbarkeit. Messtheoretische Reflexionen zur Verwendung measurement-invariance-abgesicherter Skalen in quantitativ-länderübergreifenden Settings

Tertium comparationis 24 (2018) 1, S. 57-83

Tertium Comparationis
Journal für International und Interkulturell
Vergleichende Erziehungswissenschaft
Vol. 24, No.1, pp. 57–83, 2018

Konstruktion von Vergleichbarkeit. Messtheoretische Reflexionen zur Verwendung measurement-invariance-abgesicherter Skalen in quantitativ-länderübergreifenden Settings

Jonas Scharfenberg,[1] Manuela Keller-Schneider,[2]
Sabine Weiß,[1] Meeri Hellsten,[3] Ewald Kiel[1]

[1]Ludwig-Maximilians-Universität München
[2]Pädagogische Hochschule Zürich
[3]Stockholm University

Abstract

This paper presents a methodological solution to a main challenge quantitative research has to meet: The construction of measurement equivalence. Using an international research project investigating future teachers' career choice motives, the paper demonstrates a validation approach for scales obtained from confirmatory factor analysis across countries. Measurement invariance analyses provide information on whether cross-country differences result from cultural bias, misconceptions or mistranslations during the construction of the instrument. Thus, the methodological approach helps to ensure the validity of quantitative transnational research by incorporating the demand for comparability into the instrument. The results show country specific differences among future teachers' career choice motives that can be linked to different political and social framework conditions, including the appreciation of altruistic motives, working conditions or the universities' content structure of their teacher training programs.

1. Vergleich und Vergleichbarkeit

Im Rahmen der International Vergleichenden Erziehungswissenschaft werden Fragen der Vergleichbarkeit unterschiedlicher kultureller bzw. nationaler Kontexte mittels quantitativer Methoden immer wieder kontrovers diskutiert (Blömeke, 2011; Pereyra, Kotthoff & Cowen, 2012). Diese Diskussionen erstrecken sich von der korrekten Übersetzung der eingesetzten Instrumente (z.B. Alm, Jungert &

Thornberg, 2014) über Herausforderungen durch unterschiedliche nationale Forschungstraditionen und nur begrenzt zu vereinheitlichende Begriffsverständnisse bis hin zu sehr grundlegenden Fragen wie der, „ob und was in einer international vergleichenden Betrachtung … zu lernen sei“ (Larcher & Oelkers, 2004, S. 128).

Diese Fragen stellen sich auch im Bereich der Lehrerbildung, die aufgrund der in vielen Ländern staatlich regulierten Ausbildungsstrukturen in besonderem Maße von nationalen Rahmenbedingungen beeinflusst ist (Larcher & Oelkers, 2004). Da ein Vergleich jedoch „nur dann sinnvoll angenommen werden kann, wenn die *Vergleichbarkeit* der Gegenstände schon festgestellt ist“ und Vergleichbarkeit nicht inhärent vorhanden ist, sondern erst „her-, bereit- und sichergestellt werden“ muss (Gutmann & Rathgeber, 2011, S. 49), kommt der Frage, wie man die Gleichwertigkeit des Erfassten in den einzelnen Ländern auch empirisch sicherstellen kann, besondere Bedeutung zu (Blömeke, 2011): Die Frage nach den Methoden des Vergleichs ist untrennbar verbunden mit der nach der Vergleichbarkeit der Untersuchungsgegenstände.

Die gegenwärtige Forschungslandschaft weist eine Fülle an quantitativen Methoden auf, die Gruppen- und damit Ländervergleiche zulassen. Neben Ansätzen aus dem Bereich der *Item-Response-Theory*, loglinearen Modellen sowie explorativem und varianzanalytischem Vorgehen (Berry, Poortinga, Segall & Dasen, 2002) sind hier insbesondere deduktive Verfahren aus der Kategorie der Strukturgleichungsmodelle zu nennen, etwa konfirmatorische Faktorenanalysen. Diese ermöglichen es im Gegensatz zu induktiven Verfahren, die Passung eines vorab festgelegten Modells auf spezifische Daten mittels globaler Gütekriterien zu überprüfen. Fragen der Vergleichbarkeit können konfirmatorische Faktorenanalysen in ihrer konventionellen Form jedoch nicht direkt adressieren, obwohl solche Fragen insbesondere jene Art von Daten betreffen, die die Grundlage für viele quantitative Forschungsprojekte bilden – Daten, die über Selbstauskünfte erfasst werden und damit für kulturell unterschiedliche Antworttendenzen sowie kulturspezifische Unterschiede besonders anfällig sind.

Vor diesem Hintergrund diskutiert dieser Beitrag messtheoretische Handlungsoptionen, um einer zentralen Herausforderung quantitativer Forschung zu begegnen: Der Herstellung von Messäquivalenz in unterschiedlichen Kontexten. Dazu wird mit der *Measurement-Invariance*-Messung eine Möglichkeit vorgestellt, Skalen im Rahmen von konfirmatorischen Faktorenanalysen länderübergreifend abzusichern. Derartige Analysen finden u.a. im Rahmen der international vergleichenden Psychologie (Byrne & Watkins, 2016; Milfont & Fischer, 2010) sowie bei Untersuchungen im Umfeld der internationalen Schul- (z.B. Schulz-Heidorf & Solheim, 2016) bzw. Lehrervergleichsstudien (z.B. Blömeke, 2011) Anwendung. Sie erlauben es festzustellen, ob sich die Passung des Messsystems zwischen ein-

zelnen Untersuchungsländern unterscheidet und ermöglichen plausible Aussagen darüber, ob beobachtete Differenzen zwischen Personen aus verschiedenen Ländern an der Länderzugehörigkeit selbst liegen oder ob sie durch einen *Cultural Bias* hervorgerufen werden, also beispielsweise die Folge von Übersetzungsfehlern oder länderspezifischen Unterschieden in der Operationalisierung der untersuchten Konstrukte sind (Berry et al., 2002). Eine so ergänzte konfirmatorische Faktorenanalyse ermöglicht zugleich valide Vergleiche, und eine Überprüfung der Vergleichbarkeit und kann so entscheidend zur Validität international-vergleichender Forschung beitragen.

Dieses Vorgehen wird am Beispiel eines Forschungsvorhabens aus dem Projekt ‚Student Teachers' Motives' (STeaM) vorgestellt. STeaM erhebt als Kooperationsprojekt unter Leitung der LMU München und der PH Zürich in Zusammenarbeit mit Kooperationspartnern aus weiteren Ländern die Studien- und Berufswahlmotive von Lehramtsstudierenden verschiedener Fächer, Schularten und Länder. Es dient der Ermittlung von Einflussfaktoren auf die Studien- und Berufswahlmotive, ihrem nationalen und internationalen Vergleich und der Entwicklung von Beratungsinstrumenten (Weiß & Kiel, 2013).

Bei der Ausdehnung des Projektkontexts über deutsche Studierende hinaus stellt sich die Frage, inwiefern ein für diese Personengruppe konstruierter Fragebogen auch im international-vergleichenden Kontext Anwendung finden kann. Der Übergang von einem nationalen zu einem internationalen Forschungskontext hat einerseits Implikationen für den Forschenden selbst, der sich mit Fragen der Nostri- bzw. Altrifizierung, den machttheoretischen Implikationen seiner Forschung sowie dem potenziellen Wechsel von einer Insider- zu einer Outsider-Position auseinandersetzen muss (Parreira do Amaral, 2015). Er erfordert zum anderen eine Reflexion der Kategorien des Nationalen bzw. des Kulturellen, die als Vergleichsmaßstäbe herangezogen werden und mit der Gefahr einer Überbetonung des nationalen Kontexts bzw. einer Verallgemeinerung staatlicher Vergleichsstrukturen im Sinne eines *methodological nationalism* bzw. *statism* einhergehen (Dale & Robertson, 2009). So ermöglicht länderübergreifende Forschung nicht nur eine Außensicht und damit eine vertiefte Einsicht in die Kontexte der eigenen Tätigkeit (Blömeke, 2011) und kann gegenseitige Reflexionen und Lernprozesse befruchten (s.a. Steiner-Khamsi, 2002; Larcher & Oelkers, 2004). Sie zwingt den Forschenden auch in besonderem Maße zu einer Reflexion seiner Prämissen sowie zu einer Positionierung gegenüber dem Forschungsprozess, seinem Gegenstand sowie gegenüber den Beforschten selbst.

Dem folgend wird im vorliegenden Beitrag überprüft, inwiefern das STeaM-Instrumentarium geeignet ist, über messinvariante Skalen die Motive deutscher, schwedischer und Schweizer Studierender länderübergreifend vergleichbar zu

erfassen. Das Vorgehen wird exemplarisch anhand von zwei Motivkategorien dargestellt, die aus der bisherigen nationalen und internationalen Forschung extrahiert wurden. Vor diesem Hintergrund wird im Folgenden zuerst die bisherige Forschung zu den Studien- und Berufswahlmotiven zusammengefasst, bevor die *Measurement-Invariance*-Analyse im Rahmen des Methodenteils eingeführt wird.

2. Theoretischer Hintergrund

Zur allgemeinen Studien- und Berufswahl existiert eine reiche Forschungslandschaft, an die sich die Forschung zu den Studien- und Berufswahlmotiven von Lehramtsstudierenden allerdings bis auf wenige Ausnahmen (z.B. Kaub, Stoll, Biermann, Spinath & Brünken, 2014) zumeist nur lose anlehnt. Die meisten Bezüge gibt es zu Erwartung-x-Wert-Theorien, etwa bei Watt und Richardson (2007) sowie zu *Person-Environment-Fit*-Modellen (z.B. Holland, 1997), die die Berufswahlentscheidung als Passung zwischen personalen Merkmalen, etwa dem individuellen Persönlichkeitstyp, und den Anforderungen, aber auch den Möglichkeiten eines spezifischen Berufes beschreiben.

2.1 Forschung zu Studien- und Berufswahlmotiven von Lehramtsstudierenden

Im deutschen bzw. englischen Sprachraum existiert eine breite Forschungslandschaft zu den Studien- und Berufswahlmotiven von Lehramtsstudierenden (Rothland, 2014; Zumwalt & Craig, 2008), während es im schwedischsprachigen Raum erst seit wenigen Jahren entsprechende Forschung gibt (Alm et al., 2014). Obwohl die Vergleichbarkeit der Forschungsansätze aufgrund verschiedener methodischer Verortungen, Motivinventare und zum Teil mit den Analysen verbundener normativer Werturteilen massiv eingeschränkt ist (Rothland, 2014), lassen sich doch einige häufig wiederkehrende Motivkategorien beschreiben.

Viele Arbeiten unterscheiden als Grundkategorien intrinsische und extrinsische Motive (z.B. Bastick, 2000; Kyriacou, Kunc, Stephens & Hultgren, 2003). Unter intrinsischen Motiven werden zumeist solche verstanden, die direkt mit den Tätigkeiten zukünftiger Lehrkräfte verknüpft sind, etwa Freude am Umgang mit Kindern/Jugendlichen oder Freude am Unterrichten. Extrinsische Motive zielen demgegenüber auf einen außerhalb der Tätigkeit liegenden Zweck, etwa ein hohes Gehalt oder berufliche Sicherheit. Bis auf wenige Ausnahmen (etwa bei Bastick, 2000) überwiegt die Bedeutung der intrinsischen Motive, auch bei offenem Fragedesign (Ulich, 2004).

Einige Studien ergänzen diese Grundkategorien, etwa um altruistische oder pragmatische Motive. So unterscheiden etwa Kyriacou et al. (2003) intrinsische, auf die Freude an lehrerspezifischen Tätigkeiten ausgerichtete Motive von altruisti-

schen Motiven, die den Nutzen dieser Tätigkeit für Dritte betonen. Pragmatische Motive weisen hingegen weder eine Verbindung zu den eigentlichen Tätigkeiten noch zu den extrinsischen Arbeitsbedingungen von Lehrkräften auf, sondern erscheinen weitgehend losgelöst vom Berufsbild – etwa, wenn man seine Studienwahl ohne großes Interesse oder auf Empfehlung der Eltern hin trifft. Da diese Motive mehr auf die Studien- als auf die Berufswahl fokussieren, stellen sie tendenziell eher Studien- als Berufswahlmotive dar. Die Vernachlässigung gerade der pragmatischen Motive wurde in der Forschung wiederholt kritisiert (Rauin & Römer, 2010).

Dass sich die Mehrzahl der bisher genannten Studien je auf eine einzige Untersuchungsregion beschränkt, ist symptomatisch für eine Untersuchungslandschaft, die nicht nur methodisch, sondern auch regional zersplittert ist. Dabei erscheint gerade im Bereich der Lehramtsausbildung, die als stark staatlich reglementierter Bereich in besonderem Maße staatsräumlich variierenden Rahmenbedingungen ausgesetzt ist (vgl. Larcher & Oelkers, 2004), ein Vergleich anhand räumlicher Kategorien sinnvoll, um beispielsweise die Auswirkungen unterschiedlicher Kontextfaktoren in unterschiedlichen Ländern miteinander vergleichen zu können. Dass das Nationale als Kategorie des Vergleichs dabei nicht im Sinne eines *methodological statism* übermäßig generalisiert werden darf, erschließt sich bereits aus den unterschiedlichen Strukturen innerhalb der einzelnen Untersuchungsländer (Bartlett, 2016; Dale & Robertson, 2009). So ist die Lehramtsausbildung in einigen Ländern zentralstaatlich, in anderen bundesstaatlich und in anderen wiederum unter Beteiligung lokaler Akteure strukturiert (European Commission, 2017). Dennoch unterscheidet sich der Katalog der als relevant betrachteten Motive zwischen den Forschungslandschaften der einzelnen Länder ebenso wie ihre Bewertung: Während etwa altruistisch-idealistische Motive in der deutschsprachigen Forschung teilweise negativ betrachtet und mit der Gefahr der Selbstüberlastung in Verbindung gebracht werden, gelten sie in der angloamerikanischen Forschungstradition geradezu als Grundlage einer erfüllenden Lehrertätigkeit (vgl. Schmitz & Leidl, 1999; Stiegelbauer, 1992; Kyriacou et al., 2003).

Insgesamt gibt es kaum Erkenntnisse dazu, inwiefern strukturelle Rahmenbedingungen, beispielsweise die spezifischen Bedingungen innerhalb eines Landes, die Motivlage beeinflussen (Keller-Schneider, Weiß & Kiel, im Druck). Allerdings ist diese Frage durchaus relevant: Fasst man die Studien- und Berufswahl entsprechend der *Person-Environment-Fit*-Modelle als Herstellung einer größtmöglichen Passung zwischen den Interessen eines Individuums und den Anforderungen und Angeboten eines Berufs auf, ist zu erwarten, dass strukturelle Bedingungen über eine Beeinflussung der ‚Angebotsseite' auch auf die Studienwahlmotive wirken, also beeinflussen, wer sich aus welchen Gründen für den Lehrerberuf entscheidet.

Während es aufgrund der obigen Befunde naheliegend ist, einen maßgeblichen Einfluss der strukturellen Rahmenbedingungen anzunehmen, sind die eigentlichen Wirkzusammenhänge noch weitgehend unbekannt.

2.2 Fragestellung

Nur wenige Studien haben bisher Studien- und Berufswahlmotive zukünftiger Lehrkräfte auf einer internationalen Basis untersucht (z.B. Kyriacou et al., 2003; Syring, Weiß, Keller-Schneider, Hellsten & Kiel, 2017). Erst in den letzten Jahren entstand mit *FIT-Choice* ein Modell, das länderübergreifend Anwendung fand (Watt & Richardson, 2007; für den deutschsprachigen Raum König, Rothland, Darge, Lünnemann & Tachtsoglou, 2013). Allerdings ist die direkte Vergleichbarkeit dieser Erhebungen teilweise eingeschränkt, da für die Einzelerhebungen mehrfach Skalen des Modells modifiziert, eliminiert oder ergänzt wurden (z.B. Suryani, Watt & Richardson, 2016). Gerade die fehlende Vergleichbarkeit der Ergebnisse aus der Vielzahl der bisher vorliegenden Einzelstudien zu den Studien- und Berufswahlmotiven bildet einen wesentlichen Kritikpunkt an der bisherigen Forschungslandschaft (vgl. Rothland, 2014).

Diese Untersuchung geht daher einen anderen Weg: Als Teil des STeaM-Projekts, in dessen Rahmen bereits die Studien- und Berufswahlmotive von Lehramtsstudierenden verschiedener Fächer, Schularten und Länder (z.B. Weiß & Kiel, 2013) in explorativ-quantitativen Designs untersucht wurden, geht sie den folgenden Fragen nach:

1) Inwiefern lässt sich ein Instrument konstruieren, das eine messinvariante Messung intrinsischer und pragmatischer Studien- und Berufswahlmotive erlaubt und für die Länder Deutschland, Schweiz und Schweden Gültigkeit hat?

Und aufbauend darauf:

2) In welchen Studien- und Berufswahlmotiven zeigen sich Unterschiede zwischen den Studierenden der beteiligten Länder?

Zur Beantwortung der Forschungsfragen werden Daten von deutschen, schwedischen und Schweizer Studierenden exemplarisch im Hinblick auf intrinsische und pragmatische Motive untersucht. Dies hat in erster Linie methodische Gründe, da *Measurement-Invariance*-Analysen höhere Anforderungen an den Datenumfang stellen als konventionelle konfirmatorische Faktorenanalysen und für diese Länder ausreichend Daten vorliegen, um die Analysen unmodifiziert durchführen zu können. Daneben eignen sich die drei Länder für eine erste Überprüfung der Übertragbarkeit des Modells, da sie einerseits als der OECD zugehörige Länder einer verhältnismäßig homogenen Ländergruppe entstammen, sich andererseits hinsichtlich

ihrer strukturellen Bedingungen jedoch auch unterscheiden (European Commission, 2017). So verdient etwa eine deutsche Lehrkraft 120,9 bis 198,7 Prozent des jeweiligen nationalen Durchschnittseinkommens, eine schwedische Lehrkraft hingegen nur 67,8 bis 109,6 Prozent (European Commission, EACEA & Eurydice, 2016). Ein weiteres Unterscheidungskriterium ist die Arbeitszeit. Während schwedische Lehrkräfte mindestens 35 Stunden pro Schulwoche im Schulgebäude verbringen müssen, sind deutsche Lehrkräfte in ihrer Tagesgestaltung deutlich freier, da ihnen zumeist nur ein Lehrpensum statt einer fixen Arbeitszeit vorgeschrieben wird (Eurydice, 2017a, 2017b). In der Schweiz sind die Regelungen kommunal und auch schulspezifisch, wobei immer mehr Lehrpersonen rund 45 Stunden pro Woche in der Schule verbringen. Daneben können auch Unterschiede in der Studiengestaltung relevant werden. So errechnen sich etwa bis zu 83,5 Prozent der Studienleistungen im bayerischen Lehramtsstudium aus fachwissenschaftlichen Studienanteilen, während dieser Anteil an einer der schwedischen Partnerhochschulen lediglich maximal 72,8 Prozent beträgt (LMU München, o.J.; Södertörn Högskola, o.J.). In der Ausbildung Schweizer Lehrpersonen der Primarstufe (einphasige Ausbildung) umfassen fachdidaktische, erziehungswissenschaftliche und berufspraktische Anteile demgegenüber je rund ein Drittel der Ausbildungszeit (Pädagogische Hochschule Zürich, o.J.). Da sich das deutsche Sample auf bayerische Studierende und die Schweizer Stichprobe auf Studierende aus dem Kanton Zürich beschränkt, ist die landesweite Generalisierbarkeit dieser Ergebnisse durch Unterschiede in den Rahmenbedingungen zwischen den einzelnen Kantonen bzw. Bundesländern allerdings stark eingeschränkt.

3. Methodisches Vorgehen

3.1 Projektzusammenhang und Stichprobe

STeaM baut auf einem quantitativen Forschungsdesign auf, das in einigen Untersuchungsländern im Rahmen einer Erweiterungsstudie um problemzentrierte Interviews mit Studierenden und Mitarbeiterinnen und Mitarbeitern aus der universitären Lehramtsausbildung ergänzt wird. Der aus den oben genannten Gründen für diese Untersuchung gewählte Teildatensatz umfasst insgesamt $n = 2.069$ Datensätze deutscher ($n = 1.299$), schwedischer ($n = 319$) und Schweizer ($n = 451$) Lehramtsstudierender im ersten Studiendrittel, die an einer deutschen, einer Schweizerischen sowie zwei schwedischen Hochschulen erhoben wurden. Der Datensatz ist um Studierende bereinigt, die mehr als 5 Prozent aller Fragen nicht beantwortet haben. Die unterschiedlichen Teilnehmendenzahlen erklären sich unter anderem aus der unterschiedlichen Größe der untersuchten Lehramtsstudiengänge. 76,8 Pro-

zent bzw. 1.588 der Befragten sind weiblich, das Durchschnittsalter beträgt 22,4 Jahre ($SD = 4{,}98$).

3.2 Messinstrument

Der STeaM-Fragebogen besteht aus 73 vierstufigen Likert-skalierten Items. Er wurde an der LMU München entwickelt und fand in einer Vielzahl nationaler und internationaler Studien Verwendung (u.a. Keller-Schneider, 2011; Weiß & Kiel, 2013; Syring et al., 2017; Keller-Schneider et al., im Druck). Er basiert auf den zu diesem Zeitpunkt vorliegenden empirischen Befunden aus dem Bereich der Studien- und Berufswahl von Lehramtsstudierenden sowie auf Befragungen von Expertinnen und Experten, in deren Rahmen weitere, fehlende Aspekte ergänzt wurden. Für die nicht deutschsprachigen Untersuchungsländer wurde der Fragebogen in Kooperation mit den jeweiligen Partnern in mehrstufigen, individuell abgesprochenen Verfahren in die jeweiligen Landessprachen übersetzt.

Bisher wurde der Fragebogen vor allem mittels induktiver Methoden ausgewertet. Um jedoch einen Vergleich verschiedener Länder mit Hilfe eines Motivinventars zu ermöglichen, das auch für zukünftige Datenerhebungen in Betracht gezogen und bei Veränderungen der Datengrundlage erneut auf seine Gültigkeit hin überprüft werden kann, wird für diese Untersuchung ein deduktives Vorgehen gewählt, das auf den bisherigen Erkenntnissen der deutschsprachigen und internationalen Forschung zu den Studien- und Berufswahlmotiven aufbaut.

3.3 Quantitativ-länderübergreifende Analyse von Fragebogendaten

Entsprechend der Grunderkenntnis des kritischen Rationalismus ist eine induktive Verifikation einer Theorie formallogisch nicht möglich (Popper, 1935). Stattdessen haben sich – durchaus auch in Abgrenzung zu Poppers ursprünglichen Ideen – in den quantitativ arbeitenden Sozialwissenschaften deduktive Verfahren durchgesetzt, die probabilistische Aussagen darüber zulassen, wie unwahrscheinlich das Nichtzutreffen einer Theorie ist: Eine Theorie ist zu akzeptieren, wenn die empirischen Ergebnisse unter Annahme alternativer Prämissen hinreichend unwahrscheinlich zu Stande kommen würden.

Diese Unwahrscheinlichkeit ist jedoch wiederum selbst Element wissenschaftlicher Diskussion, da sie ebenfalls nur geschätzt werden kann und ihre Grenzwerte letztendlich normative Setzungen sind. Unterschiedliche Schätzverfahren, unklare Grenzwerte und divergierende Generalisierungsansprüche können zu Unterschieden hinsichtlich der Frage führen, was in welchem Umfang wann als hinreichend unwahrscheinlich gilt, um eine Theorie zu stützen. Zugleich kann sich auch die

Auffassung davon ändern, was als ausreichend gilt, um eine Theorie als plausibel anzunehmen.

Verwendet man beispielsweise konfirmatorische Faktorenanalysen zur Datenanalyse, überprüft man in einer Reihe von Tests die unterschiedlichen Elemente des Modells auf Plausibilität bzw. Unwahrscheinlichkeit ihres zufälligen Auftretens. Die Angaben dazu, welche Tests durchzuführen und welche Grenzwerte einzuhalten sind, variieren jedoch abhängig vom Autor bzw. von der Autorin und wurden im Laufe der Zeit teilweise verschärft (z.B. Backhaus, Erichson & Weiber, 2015; Schermelleh-Engel, Moosbrugger & Müller, 2003). Im Folgenden wird daher zuerst dargestellt, welche Tests die Validität konventioneller konfirmatorischer Faktorenanalysen sicherstellen und diskutiert, inwiefern diese Maßnahmen im Kontext international-vergleichender Forschung ausreichen. Im Anschluss wird mit der *Measurement-Invariance*-Analyse ein Verfahren vorgestellt, das neben der Qualität des Vergleichs als solchem auch die Vergleichbarkeit der Daten thematisieren kann.

3.3.1 Die konfirmatorische Faktorenanalyse als deduktives Verfahren

Bei Fragebögen ist es Praxis, ein Konstrukt (in diesem Fall ein Motiv) nicht nur über eine einzige Frage, sondern mittels eines Messmodells aus mehreren Fragen zu erheben. Damit wird dem Umstand Rechnung getragen, dass sich die subjektive Bedeutung, die Individuen einzelnen Wörtern, Formulierungen etc. beimessen, nicht kontrollieren lässt und das individuelle Verständnis einer Frage von Person zu Person wechselt (Backhaus et al., 2015). Da das Verständnis jeder Einzelfrage individuell großen Schwankungen unterworfen ist, wird davon ausgegangen, dass Fragen, die trotz dieser individuellen Schwankungen regelmäßig und von vielen Befragten in hohem Maße ähnlich beantwortet werden, auf ein gemeinsames, interindividuell stabiles Konstrukt hinweisen.

Konfirmatorische Faktorenanalysen werden als strukturprüfende Verfahren zur Kontrolle solcher Annahmen eingesetzt (für technische Informationen vgl. Jöreskog, 1969), indem angenommen wird, dass die Antworten mehrerer Fragebogenfragen als manifeste Variablen x_i einem Konstrukt, d.h. einer latenten, nicht direkt messbaren Variable ξ_j, zugeordnet werden können. Dabei wird die latente Variable als Faktor bezeichnet, die Beziehungen der einzelnen Fragen auf diesen Faktor als Ladungen. Hohe Ladungen, d.h. starke Korrelationen einer Frage zu einer latenten Variable und niedrige Ladungen gegenüber den anderen latenten Variablen garantieren, dass das Verhältnis zwischen einer Frage und der dahinterliegenden latenten Variable belastbar ist.

Diese Belastbarkeit wiederum lässt sich mit Hilfe verschiedener Kennzahlen ausdrücken, die jede Ebene des Modells umfassen – angefangen von der einzelnen

Frage über die einzelnen Konstrukte bis hin zum Gesamtmodell. In einem ersten Schritt wird auf Ebene der Einzelzuordnung zwischen Frage und Faktor untersucht, ob die jeweilige Ladung signifikant ist. So wird sichergestellt, dass einzelne Fragen nicht Konstrukten zugeordnet werden, denen sie nicht zugehörig sind. Auf Konstruktebene wird zumeist eine Analyse der Faktorreliabilität vorgenommen, die die innere Konsistenz der Faktoren misst (Backhaus et al., 2015). Zuletzt testen verschiedene *Global-Fit*-Parameter die Güte des Gesamtmodells. Hier ist insbesondere zwischen *Goodness-of-Fit*-Parametern zu unterscheiden, die das vorgeschlagene Modell mit einem perfekt an den Datensatz angepassten Modell vergleichen, und *Badness-of-Fit*-Indizes, die das Modell mit einem Basismodell vergleichen und die die Verbesserung gegenüber diesem schlechten Modell messen (Schermelleh-Engel et al., 2003).

Für dieses Modell wird der Modellfit mittels eines *Goodness-of-Fit*-Parameters (RMSEA), mittels eines *Badness-of-Fit*-Parameters (SRMR) sowie des *Comparative Fit Index* (CFI) überprüft. Der CFI ist zwar weit verbreitet, benachteiligt allerdings im Gegensatz zum RMSEA komplexe Modelle, weswegen dem RMSEA im Zweifelsfall das größere Gewicht eingeräumt wird (Cheung & Rensvold, 2002). Die Berechnungen der Strukturgleichungsmodelle können mit Hilfe verschiedener Softwarelösungen erfolgen; diese Untersuchung benutzt die Software R (R Core Team, 2016) mit den Packages lavaan (Rosseel, 2012) und semTools (semTools Contributors, 2016). Zur Identifizierung des Modells werden alle Faktorvarianzen auf 1 festgesetzt.

3.3.2 *Measurement-Invariance*-Analysen in länderübergreifenden quantitativen Forschungsprojekten

Die bisher beschriebenen Techniken reichen aus, um einen Gesamtdatensatz zu überprüfen und ein Modell bei Nacherhebungen auf die neuen Daten anzuwenden, falls es auch dort die Gütekriterien erfüllt. Somit weist es bereits eine deutlich höhere Stabilität und Anwendbarkeit als explorativ konstruierte Modelle auf. Um Daten unterschiedlicher Länder miteinander zu vergleichen, reicht dieses Vorgehen jedoch noch nicht aus, denn bisher überprüft das Modell lediglich den Gesamtdatensatz mit den Daten aus allen darin enthaltenen Ländern, was noch nichts über die länderspezifische Passung aussagt: Es ermöglicht noch keine Aussagen zur Vergleichbarkeit des Gemessenen in den verschiedenen Untersuchungsländern. Gerade die nachträgliche Übertragung eines Instruments geht jedoch immer mit der Möglichkeit eines *Cultural Bias* einher. Dieser kann entweder auf Ebene eines Einzelitems auftreten (z.B. durch Übersetzungsfehler), als Operationalisierungs- bzw. Konzeptbias einzelne Konstrukte betreffen (z.B. eine länderübergreifend unvollständige Operationalisierung) oder als Methodenbias (z.B. durch unterschiedli-

che Skaleninterpretationen) die gesamte Erhebung modifizieren (Berry et al., 2002; Byrne & Watkins, 2016). Dies kann so weit gehen, dass ein untersuchtes Konstrukt in einzelnen Untersuchungskontexten gar nicht existent ist. Allerdings gehen mit diesen Formen des *Cultural Bias* spezifische Veränderungen einher, die sich an der Datenstruktur zeigen lassen.

So könnte etwa eine konfirmatorische Faktorenanalyse eine Ladung $\lambda_{ij} = 0{,}60$ einer Variable x_i auf einen Faktor ξ_j messen. Geht man davon aus, dass der Datensatz aus zwei Länderdatensätzen besteht, könnte es durchaus sein, dass die Ladung in beiden Ländern annähernd identisch ist – etwa $\lambda_{ijA} = 0{,}59$ in Land A und $\lambda_{ijB} = 0{,}61$ in Land B. In diesem Fall ist eine länderübergreifende Messung möglich, denn es gilt hier, was schon bei der Fragebogenkonstruktion galt: Wenn trotz aller interindividuellen Unterschiede mehrere Fragen tendenziell gleich beantwortet werden und sich dieses Verhältnis selbst zwischen Angehörigen unterschiedlicher Gruppen kaum unterscheidet, kann man davon ausgehen, ein stabiles Konstrukt zu messen. Die Bezeichnung dieses Konstrukts ist eine hermeneutische Aufgabe, doch die Messung an sich lässt sich plausibilisieren: Würden sich Bedeutungszuschreibungen oder das Verständnis eines Konzepts zu sehr verändern, erscheint es hochgradig unwahrscheinlich, dass diese Veränderung alle Fragen gleichermaßen betrifft und sich Konstrukte, die sich aus dem Verhältnis mehrerer Fragen zueinander zusammensetzen, nicht verändern.

Neben der Möglichkeit sehr ähnlicher Ladungen in beiden Ländern könnte die Ladung aus den oben genannten Gründen auch in einem der beiden Länder sehr stark ausgeprägt sein, während sie in dem anderen Land kaum vorhanden ist. Im Falle eines Bias, der nur einzelne Items betrifft, wird eine solche Abweichung tendenziell auch nur einzelne Ladungen betreffen. Wenn der *Cultural Bias* eines der gemessenen Konstrukte betrifft, sodass es beispielsweise nur noch von einem Teil der zugeordneten Fragen erfasst wird, werden tendenziell mehrere Fragen des bzw. der betroffenen Konstrukte divergierende Ladungen aufweisen. Allerdings kann diese Art des Bias in einigen Fällen auch unentdeckt bleiben, insbesondere dann, wenn ein Konstrukt in einem anderen Untersuchungskontext über weitere Elemente verfügt, die durch das Messinstrument nicht abgedeckt werden (Berry et al., 2002). Diejenigen Fälle des methodischen Bias, die zu einer linearen Verschiebung sämtlicher Itemausprägungen zwischen den Gruppen führen (z.B. ein durchgehend positiveres Antwortverhalten), können sich schließlich als Achsenabschnittsverschiebung der Faktorladungs-Regressionsgeraden bei sämtlichen Ladungen zeigen (Brown, 2015). Helfrich (2013) benennt neben der Skalenäquivalenz insgesamt drei weitere validitätsbezogene Äquivalenzbegriffe, die für eine länderübergreifende Vergleichbarkeit quantitativer Skalen erfüllt sein müssen: Neben der funktionalen und der operationalen Äquivalenz, d.h. vergleichbaren Beziehungen zwischen den

einzelnen Fragen und dem Gesamtkonstrukt in allen Untersuchungsländern, insbesondere auch die konzeptuelle Äquivalenz.

Gerade im Hinblick auf solche Fragen stellt die Kontrolle der *Measurement Invariance* einen sinnvollen Zusatz für konventionelle konfirmatorische Faktorenanalysen dar: Das Verfahren ermöglicht es zu überprüfen, inwiefern sich die Messmodelle beliebig definierter Gruppen innerhalb der Daten unterscheiden (für technische Details vgl. Meredith, 1993) und gibt so konfirmatorischen Faktorenanalysen die „capability to examine the equivalence of all measurement and structural parameters of the factor model across multiple groups" (Brown, 2015, S. 241). Da die oben beschriebenen Bias zu Divergenzen in den *Measurement-Invariance*-Werten führen, erlaubt sie es insbesondere zu unterscheiden, ob beobachtete Differenzen in den Gruppen selbst liegen oder im Instrument, das benutzt wird: Erzielt das Motiv A bei Befragten des Landes B eine höhere Ausprägung als in den anderen Ländern, weil es dort relevanter ist, oder deswegen, weil das Instrument dort aus einem der oben genannten Gründe anders misst? Im ersten Fall, bei einem tatsächlich vorhandenen Gruppenunterschied, blieben die Werte für die *Measurement Invariance* unauffällig. Im zweiten Fall würden sich aufgrund der oben beschriebenen Ladungsunterschiede zwischen den Gruppen signifikante Unterschiede zeigen. Durch iterative Analysetechniken (z.B. Jöreskog, 1969) kann festgestellt werden, welche Modellrestriktionen, d.h. welche Zuordnungen, für die gruppenspezifischen Abweichungen sorgen und ob diese Abweichungen durch unverbundene Einzelitems, z.B. aufgrund fehlerhafter Übersetzungen, oder aufgrund von mehreren Items innerhalb eines Konstrukts hervorgerufen werden. Ein Lösen dieser falschen Zuordnungen führt zu einer erneuten Konvergenz der unterschiedlichen Gruppenlösungen und damit zu einem Absinken des Kennwerts in den unkritischen Bereich.

Aus messtheoretischer Sicht sichert ein solches Vorgehen parallel zum Vergleich die Vergleichbarkeit des Verglichenen, indem es methodeninhärente Kennwerte zur operationalen bzw. konzeptuellen Äquivalenz des Gemessenen zwischen den Ländern bereitstellt, die Aussagen darüber zulassen, ob gruppenspezifische Abweichungen so unwahrscheinlich sind, dass von einer validen länderübergreifenden Messung gesprochen werden kann. Um diese Plausibilität zu gewährleisten, muss eine Invarianz auf mindestens drei aufeinander aufbauenden Ebenen nachgewiesen werden (Beaujean, 2014; Brown, 2015), wobei die Modellrestriktionen mit jeder Ebene zunehmen.

Die Basis bildet eine Stufe, auf der nur die Modellstruktur, also die Zuordnung der manifesten zu den latenten Variablen, vorgegeben ist, alle anderen Werte jedoch innerhalb länderspezifischer Modelle unabhängig berechnet werden: So können die einzelnen Ländermodelle länderspezifische Unterschiede abbilden, deren Fixierung auf einen gemeinsamen Mittelwert in den folgenden Stufen die Grundla-

ge für die Analyse der *Measurement Invariance* bildet (1. Ebene, konfigurale Invarianz). In der Folge werden die Faktorladungen (2. Ebene, schwache Invarianz) und die Intercepts, d.h. die Achsenabschnitte der Faktorladungs-Regressionsgeraden (3. Ebene, starke Invarianz) über alle Gruppen hinweg auf ihre jeweiligen Mittelwerte fixiert und der Einfluss dieser Angleichungen auf die *Global-Fit*-Parameter der Modelle untersucht: Waren sich beide Gruppen von vornherein sehr ähnlich, wird eine Fixierung auf einen gemeinsamen Wert die Modellgüte kaum beeinträchtigen – die *Global-Fit*-Indices bleiben gleich. Werden jedoch zwei Werte, die aus einem der oben genannten Gründe sehr unterschiedlich ausfallen, auf ein gemeinsames Mittel fixiert, sinkt die Modellgüte.

Bereits die schwache Invarianz liefert eine begrenzte Aussagekraft über die Vergleichbarkeit der beiden Gruppen. Die Fixierung der Ladungsstruktur plausibilisiert eine ähnliche qualitative Struktur der Konstrukte in den einzelnen Untersuchungskontexten, kann jedoch Effekte wie eine Tendenz zu positiveren Antworten in einem der beteiligten Länder noch nicht abbilden (Brown, 2015). Solche Effekte sind für einen methodischen Bias typisch und führen dazu, dass sich die Ausprägungen der einzelnen Werte zwischen den Gruppen noch unterscheiden können.

Erst auf der Ebene der starken Invarianz ist eine vollständige Vergleichbarkeit erreicht: „If there are group differences on strongly-invariant LVs, it likely indicates that there are real differences in these variables, as opposed to the difference being in how the LVs are measured“ (Beaujean, 2014, S. 59). Erst die Fixierung der Intercepts im Rahmen der starken Invarianz sorgt dafür, dass auch absolute Werte zwischen den Gruppen vergleichbar werden – dass also eine Person mit dem gleichen Wert in der latenten Variable in allen Gruppen vergleichbare manifeste Variablen, d.h. Fragebogenantworten, aufweisen würde. Hierin liegt ein Vorteil gegenüber alternativen Methoden zur Abschätzung von Messinvarianz, die zwar vielfach ebenfalls Hinweise auf eine vergleichbare Faktorenstruktur der Konstrukte liefern, allerdings nur bedingt Hinweise auf die Vergleichbarkeit der absoluten Werte geben können (Berry et al., 2002).

In der Untersuchungspraxis ist immer eine starke Invarianz anzustreben, auch wenn dieses Ziel gerade in der *Cross-Cultural-Psychology* von einigen Forschenden als kaum zu erreichen angesehen wird (Milfont & Fischer, 2010). Daher gibt es die Möglichkeit, einzelne Variablen von der Invarianzmessung auszunehmen – etwa, wenn man weiß, dass eine bestimmte Zuordnung zwischen den Gruppen variant ist, ohne dass man sie eliminieren will. Auf dieses als *partial invariance* bezeichnete Verfahren wird hier nicht näher eingegangen (für Details s. z.B. Brown, 2015). *Measurement-Invariance*-Analysen stellen sehr hohe Anforderungen an den Umfang der Datengrundlage, da jede Einzelgruppe über genügend Fälle verfügen muss, um das komplette Strukturgleichungsmodell zu rechnen. Wo nicht genügend

Fälle vorliegen und eventuell auch nicht erhoben werden können, bietet sich ein Zusammenschluss verschiedener Gruppen an, etwa zu Ländergruppen statt Einzelländern, um zumindest in begrenztem Maße Aussagen über Gruppenunterschiede zuzulassen.

Anschließend werden die Veränderungen des *Global Fit* anhand der Veränderungen von Kennwerten wie dem CFI- oder dem RMSEA-Index betrachtet. Da auch hier die oben genannten Einschränkungen des CFI für komplexe Modelle gelten, wird für diese Untersuchung die Veränderung des RMSEA als relevanter Wert herangezogen. Veränderungen des RMSEA um weniger als 0,01 können als starkes Indiz dafür gewertet werden, dass das Instrument stabil ist und die gemessenen Unterschiede in realen Unterschieden zwischen den Gruppen liegen (Cheung & Rensvold, 2002). Aus wissenschaftstheoretischer Perspektive wird der zufällige Eintritt derartig ähnlicher Strukturen über verschiedene Gruppen hinweg als so unwahrscheinlich betrachtet, dass von einer Äquivalenz der Strukturen ausgegangen werden darf. Falls sich ein messinvariantes Messmodell nachweisen lässt, werden die Länder mittels Brown-Forsythe-korrigierter Varianzanalysen und Post-Hoc-Tests nach Games-Howell verglichen (Games & Howell, 1976). Zur Messung der Effektstärke wird Cohen's *f* mittels G*Power (Faul, Erdfelder, Lang & Buchner, 2007) aus den univariaten Analyseergebnissen des SPSS-Outputs errechnet; da hierbei keine robusten Methoden Anwendung finden können, sind sie tendenziell vorsichtig zu interpretieren.

4. Empirische Ergebnisse

Die theoretische Grundlage der Faktorenanalyse bildet eine Analyse der bisherigen deutschsprachigen und internationalen Forschungsliteratur zu den Studien- und Berufswahlmotiven von Lehramtsstudierenden. Dabei wird nach Motiven gesucht, die trotz der unterschiedlichen Erhebungsmethoden, Benennungen und Kategorisierungen regelmäßig als Studien- und Berufswahlmotive von Lehramtsstudierenden genannt werden. Im Gegensatz zu den allgemeinen Berufswahltheorien, die nur sehr breite Motive abdecken, sind so deutlich spezifischer auf den Lehramtsberuf zugeschnittene Motive zu erwarten. Dieser Artikel konzentriert sich dabei auf intrinsische und pragmatische Motive.

4.1 Konstruktion ländervergleichender Motivskalen

Nach iterativen Optimierungsprozessen (vgl. Jöreskog, 1969) zur Verbesserung der Modellgüte lassen sich in Einklang mit dem bisherigen Forschungsstand insgesamt neun intrinsische und vier pragmatische Motivskalen bilden. Zu anderen Motiven, etwa der Fachausbildung bzw. den günstigen Studienbedingungen, lassen sich hin-

gegen keine länderübergreifend stabilen Skalen bilden. Somit ergibt sich unter Verwendung eines robusten ML-Schätzers ein Faktorenmodell mit insgesamt 13 nicht orthogonalen, d.h. potentiell miteinander korrelierenden Faktoren. Tabelle 1 gibt eine Übersicht über die gebildeten Skalen.

Tabelle 1: Ausgewählte Studien- und Berufswahlmotive von Lehramtsstudierenden

Motivskala	Itemanzahl	Beispielitem: Ich habe mich für den Lehrberuf entschieden …	Faktorreliabilität (> 0,60)
Arbeit mit Kindern und Jugendlichen	5	… weil mich die soziale Arbeit mit Kindern und Jugendlichen reizt.	0,75
Zusammensein mit Kindern und Jugendlichen	3	… weil ich Kinder gerne mag.	0,71
Förderung von Kindern und Jugendlichen mit besonderen Bildungsvoraussetzungen	4	… um lernschwache Kinder/Jugendliche zu fördern.	0,75
Freude am Unterrichten	3	… weil es Spaß macht, anderen etwas beizubringen.	0,68
Fachliches Interesse	4	… weil ich mein Unterrichtsfach/meine Unterrichtsfächer für wichtig halte.	0,80
Gesellschaftliche Relevanz	2	… um die Gesellschaft zu verändern.	0,67
Anspruchsvoller Beruf	3	… um später einen herausfordernden Beruf zu haben.	0,62
Polyvalenz des Studiums	2	… weil die Studieninhalte nützlich sind, auch wenn man später nicht als Lehrer/in arbeitet.	0,64
Interesselosigkeit	3	… weil es keine Berufe gibt, die mich reizen.	0,66
Studium als Notlösung	3	… obwohl ich etwas ganz anderes studieren wollte.	0,72
Externe Einflüsse	4	… weil mir meine Mutter dazu geraten hat.	0,75

Die länderübergreifende Messung verlangt, wie oben beschrieben, eine doppelte Validierung der Skalen: Die Kontrolle, ob die konfirmatorische Faktorenanalyse als solche Skalen hervorbringt, die eine reliable und valide Messung der latenten Faktoren erlauben, wird über eine Analyse der Faktorladungen, der Faktorreliabilitäten und der *Global-Fit*-Indices erreicht. Für die oben genannten Skalen können diese Bedingungen als erfüllt angenommen werden: Alle Faktorladungen sind mit

$p < 0{,}001$ signifikant und alle Faktorreliabilitäten liegen über dem Grenzwert von 0,6. Einen Überblick über die *Global-Fit*-Parameter bietet Tabelle 2.

Tabelle 2: *Global-Fit*-Parameter des Modells der Studien- und Berufswahlmotive (Schermelleh-Engel et al., 2003)

Fit-Index	Guter Fit	Akzeptabler Fit	Erreichter Wert
RMSEA	0 ≤ RMSEA ≤ ,05	,05 ≤ RMSEA ≤ ,08	,05
SRMR	0 ≤ SRMR ≤ ,05	,05 < SRMR ≤ ,10	,04
CFI	,97 ≤ CFI ≤ 1,00	,95 ≤ CFI < ,97	,90

Es zeigt sich, dass RMSEA und SRMR auf einen insgesamt guten Fit des Modells hindeuten, während der CFI-Wert, wie aufgrund der Modellkomplexität zu erwarten, ungünstiger ausfällt. Auch er deutet jedoch auf einen zumindest akzeptablen Modellfit hin, gerade, da seine untere Grenze zum Teil auch mit 0,90 angegeben wird (Cheung & Rensvold, 2002; Hu & Bentler, 1999).

In einem zweiten Schritt ist zu überprüfen, ob das validierte Modell einen messinvarianten Vergleich der einzelnen Gruppen zulässt. Das verwendete Programm prüft hierzu die *Measurement Invariance* durch eine automatisierte Berechnung und gibt die Veränderungen der *Global-Fit*-Indices bei schwacher und starker *Measurement Invariance* entsprechend der obigen Beschreibung aus. Bei schwacher Invarianz ergibt sich ein RMSEA-Delta von < 0,001, bei starker Messinvarianz beträgt es 0,005. Da beide Werte unter dem Grenzwert von 0,01 liegen, ist von einer invarianten Messung auszugehen: Zwar kann nicht nachgewiesen werden, dass die Modellstruktur in allen drei Ländern vergleichbar ist, doch wäre eine so geringe Veränderung der Modellgüte bei unterschiedlichen Motivstrukturen so unwahrscheinlich, dass mit hinreichender Wahrscheinlichkeit von einer gruppenübergreifenden Vergleichbarkeit der Motive gesprochen werden kann. Statistisch ist die Grundlage für eine länderübergreifende Vergleichbarkeit der Studien- und Berufswahlmotive mittels der oben genannten Skalen gelegt.

4.2 Länderspezifische Motivunterschiede

Die Ergebnisse weisen auf länderspezifische Unterschiede bei allen Motiven außer der *Förderung von Kindern und Jugendlichen mit besonderen Bildungsvoraussetzungen* hin (vgl. Tab. 3). Die Mehrzahl der Post-Hoc-Tests fällt ebenfalls signifikant aus.

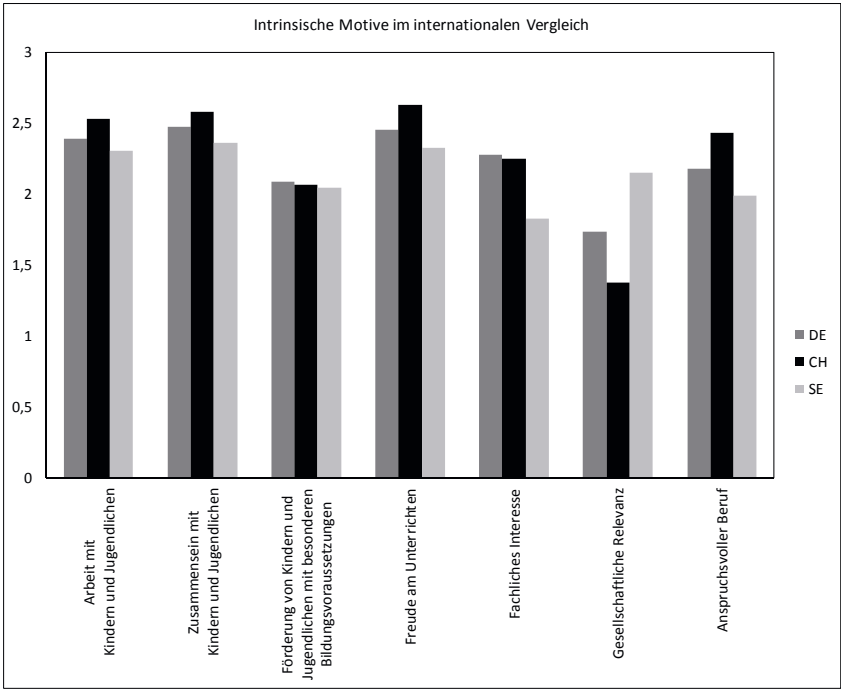
Tabelle 3: Länderspezifische Unterschiede in den Studien- und Berufswahlmotiven

Skala	ANOVA (Skala ~ Land; df1 = 2)		Cohen's	Nicht signifikante Post-Hoc-Tests ($p > 0,05$)
	df2	*F*	*f*	
Arbeit mit Kindern und Jugendlichen	981,64	27,949**	0,16	
Zusammensein mit Kindern und Jugendlichen	831,53	16,28**	0,13	
Förderung von Kindern und Jugendlichen mit besonderen Bildungsvoraussetzungen	955,63	0,635	---	---
Freude am Unterrichten	963,12	38,707**	0,2	
Fachliches Interesse	742,2	61,184**	0,27	DE - CH ($p = 0,58$)
Gesellschaftliche Relevanz	2066	108,526**	0,32	
Anspruchsvoller Beruf	843,99	56,066**	0,25	
Polyvalenz des Studiums	998,33	3,359*	0,06	Sämtliche Post-Hoc-Tests
Interesselosigkeit	854,11	57,551**	0,23	DE - SE ($p = 0,89$)
Studium als Notlösung	851,16	11,008**	0,11	DE - CH ($p = 0,09$)
Externe Einflüsse	2066	6,877*	0,08	SE - CH ($p = 0,95$)

DE = Deutschland, SE = Schweden, CH = Schweiz. Signifikanz: ** $p < 0,001$; * $p < 0,05$.

Im Bereich der intrinsischen Motive zeigen sich zwar bei sechs der sieben Motive Unterschiede auf einem Signifikanzlevel von $p < 0,001$, allerdings fallen die Effektstärken zumeist nur schwach aus und sind somit nur von begrenzter Bedeutung. Lediglich drei der Motive zeigen mittlere Effektstärken von $f \geq 0,25$: Das *fachliche Interesse*, die *gesellschaftliche Relevanz* und der *anspruchsvolle Beruf*. Insgesamt erreichen die Schweizer Studierenden auf vier der sechs signifikanten Skalen die höchsten Werte (vgl. Abb. 1), allerdings weist nur einer dieser Unterschiede eine mittlere Effektstärke auf – der *anspruchsvolle Beruf*. Die anderen beiden mittleren Effekte werden einmal durch die deutlich niedrigeren Werte schwedischer Studierender beim *fachlichen Interesse* hervorgerufen. Zum anderen entstehen sie bei der *gesellschaftlichen Relevanz*, wo schwedische Studierende die höchsten, Schweizer Studierende hingegen die niedrigsten Werte erreichen: Während kinds- und unterrichtsbezogene Werte für Schweizer Studierende einen größeren Wert einnehmen als für Studierende anderer Länder, kehrt sich dieses Verhältnis bei einem altruistischen Motiv wie der *gesellschaftlichen Relevanz* um.

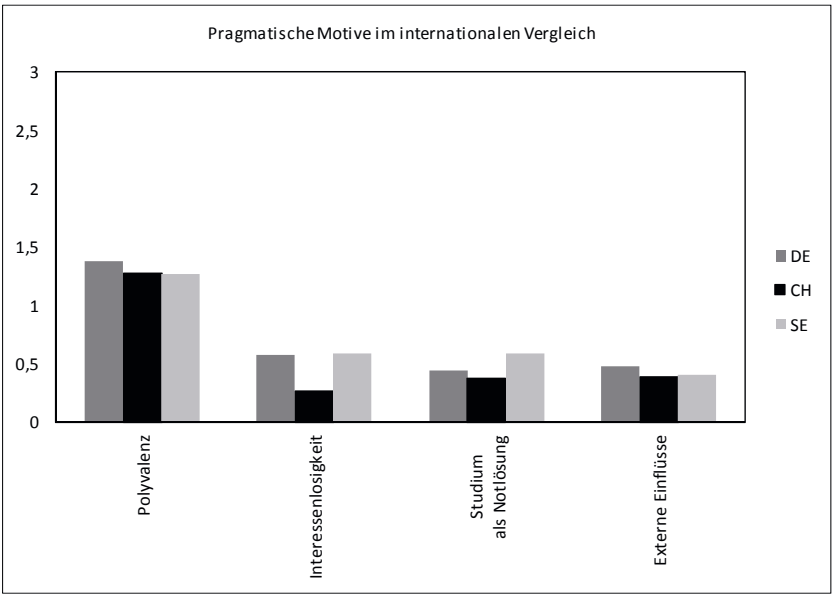
Abbildung 1: Intrinsische Studien- und Berufswahlmotive im Ländervergleich



Die Skalenmittelwerte der pragmatischen Motive fallen im Durchschnitt in allen Ländern niedriger aus als die der intrinsischen Motive (vgl. Abb. 2).

Zwar sind hier alle Skalen in Bezug auf Länderunterschiede signifikant, die Effektstärken bleiben jedoch durchgehend im kleinen bis sehr kleinen Bereich, was die Bedeutung der Unterschiede limitiert. Während Schweizer Studierende bei vier der sechs intrinsischen Motive die höchsten Werte erzielen, fallen ihre Skalenmittelwerte auf allen pragmatischen Skalen am niedrigsten aus, auch wenn die Unterschiede nur in zwei Fällen signifikant werden. Die größte Effektstärke erzielt mit $f = 0{,}23$ die Skala der Interesselosigkeit.

Abbildung 2: Pragmatische Studien- und Berufswahlmotive im Ländervergleich



5. Diskussion

Die Ergebnisse dieser Untersuchung sind sowohl auf forschungsmethodischer als auch auf inhaltlicher Ebene relevant. Daher werden zuerst die Implikationen des Ländervergleichs diskutiert, bevor die verwendete Forschungsmethodik einer Reflexion unterzogen wird.

5.1 Quantitativer Ländervergleich

Zehn der elf untersuchten Motive weisen länderspezifische Unterschiede auf. Hierbei gibt es motivübergreifende Muster: So äußern etwa Schweizer Studierende bei sechs der sieben intrinsischen Motive die höchste Zustimmung, während sie bei den pragmatischen Motiven durchgehend die niedrigsten Werte erreichen.

Einzelne länderspezifische Differenzen lassen sich mit Unterschieden in den strukturellen Bedingungen in Beziehung setzen: So könnte man etwa die in Schweden geringeren Werte beim Motiv der abwechslungsreichen Tätigkeit mit einer deutlich stärkeren Regulierung der Arbeitszeit in Verbindung bringen, die mit einer weitreichenden Standardisierung einzelner Arbeitsbereiche und langen Anwesen-

heitszeiten im Schulgebäude einhergeht. Im Gegensatz dazu spielen idealistische Motive in Schweden eine deutlich größere Rolle als in der Schweiz bzw. Deutschland, wo sie tendenziell eher kritisch gesehen werden (z.B. Schmitz & Leidl, 1999). Dies wird angehenden Lehrkräften möglicherweise schon in der Ausbildung vermittelt. Zuletzt lässt sich das fachliche Interesse der Studierenden mit den fachwissenschaftlichen Anteilen der Lehramtsausbildung in Verbindung bringen: In Deutschland, wo fachwissenschaftliche Anteile eine große Rolle spielen, ist das Motiv stärker ausgeprägt als unter schwedischen Studierenden; allerdings trifft dieser Erklärungsansatz für die Schweizer Daten nur bedingt zu, da hier die fachwissenschaftlichen Anteile am Studium deutlich geringer ausfallen als in den anderen Untersuchungsländern.

Zugleich verweist dieses Ergebnis erneut darauf, die nationale Verfasstheit der Vergleichskategorien nicht überzubewerten (vgl. Dale & Robertson, 2009): Zwar ist es plausibel, im Kontext der Lehramtsausbildung, die mehr als andere Bereiche von staatlichen Institutionen und staatlichen Vorgaben reguliert wird, staatlichen Einflüssen eine Bedeutung zuzuschreiben, doch darf darüber der Einfluss weiterer, regional, lokal, transnational oder nicht spatial erfassbarer organisierter Einflussfaktoren nicht übersehen werden. Gerade wenn ganze Länder nur über eine Hochschule erfasst werden, wie das in dieser Untersuchung für Deutschland und die Schweiz der Fall ist, ist die Unterscheidbarkeit von Phänomenen der lokalen, regionalen und nationalen Ebene sowie nicht räumlich verortbarer Phänomene stark eingeschränkt (s.a. Bartlett, 2016, S. 13). Im engen Sinn analysiert diese Untersuchung daher nicht deutsche und Schweizerische, sondern bayerische und Zürcher Rahmenbedingungen.

Da der Fokus dieses Beitrags auf dem forschungsmethodischen Vorgehen liegt, können hier derartige Zusammenhänge nur schlaglichtartig beleuchtet werden. Um sie belastbarer formulieren zu können, ist ein umfassenderer und detaillierterer Vergleich der Bedingungen und ihrer räumlichen Verortung in den Untersuchungsländern nötig, der im Idealfall um qualitative Daten ergänzt wird, die die Verbindung von individueller Berufswahl und strukturellen Einflussfaktoren beleuchten. Zugleich sind für ein vollständiges Bild auch extrinsische Motive in Betracht zu ziehen.

Ausgehend von den Ergebnissen ergeben sich drei weitere Forschungsschritte: Die Erweiterung des Messmodells um extrinsische Motive würde zu einer vollständigeren Abdeckung des Motivinventars angehender Lehrkräfte führen. Die Ausweitung des Länderkreises um Nicht-OECD-Länder würde es erlauben, die Generalisierbarkeit des Messmodells auch für Länder mit heterogeneren Bedingungen als den hier untersuchten zu überprüfen. Schließlich könnte eine Pluralisierung der Methoden und das Erheben weiterer qualitativer Daten auch für die hier untersuch-

ten Länder dazu beitragen, die Wirkmechanismen der strukturellen Einflüsse auf die Studien- und Berufswahlmotive besser zu verstehen.

5.2 Methodische und metatheoretische Diskussion

Der Fokus dieses Beitrags liegt auf der Erkenntnis, dass die Bildung messinvarianter Skalen über alle drei Länder hinweg möglich ist. Allerdings ist der untersuchte Länderkreis eingeschränkt; darüber hinaus können Motiverweiterungen aufgrund der Limitationen der Methode in einzelnen Untersuchungsländern nicht ausgeschlossen werden, während andererseits die Skalenbildung insbesondere im Bereich der pragmatischen Motive nicht für alle erwarteten Motive möglich ist. So ließ sich etwa aus den Studienbedingungen keine länderübergreifend stabile Skala bilden: Offensichtlich sind die Studienbedingungen zu unterschiedlich und variieren auf zu vielen Ebenen zwischen den Ländern, als dass man sie mit einem gemeinsamen Kennwert abfragen könnte. Zugleich zeigt sich jedoch, dass gerade intrinsische Motivlagen gut auf andere Länder übertragbar sind. Dies mag auch daran liegen, dass in Anlehnung an die unter anderem von Blömeke (2011) aufgestellten Grundbedingungen valider quantitativ-international vergleichender Forschung die konzeptuelle Vergleichbarkeit der Übersetzungen durch intensiven Austausch unter den Projektpartnern sichergestellt wurde.

Im Rahmen von *Measurement-Invariance*-Untersuchungen kann einem Grundeinwand an länderübergreifende quantitative Forschung begegnet werden: Dass Messinstrumente nicht in der Lage seien, mit kulturell bzw. länderspezifisch unterschiedlichen Konzepten und Bedeutungszuschreibungen umzugehen, dass derartige Unterschiede nur unzureichend reflektiert würden und in der Analyse der Ergebnisse nicht kenntlich gemacht würden (Cheung & Rensvold, 2002). Die epistemologische Frage, inwiefern komplexe Konzepte oder Selbstkonzeptbestandteile wie Motive überhaupt im Rahmen quantitativer Operationalisierungen messbar gemacht werden können, kann auch von dieser Methode nicht beantwortet werden. Allerdings bietet die doppelte Absicherung über eine Skalenkonstruktion, die auf der Ähnlichbeantwortung mehrerer Fragen und der länderübergreifenden Kontrolle dieser Konstrukte beruht, eine sehr sichere Basis, um von einer grundsätzlichen Äquivalenz der gemessenen Konstrukte ausgehen zu können (Brown, 2015).

Allerdings hat die *Measurement-Invariance*-Methodik auch Grenzen: So ist die Identifizierbarkeit von methodischem Bias ohne externe Referenzrahmen vor allem innerhalb der *Cross-Cultural-Psychology* umstritten (Berry et al., 2002). Dies gilt insbesondere, da viele Simulationsstudien zur Wirksamkeit der Methode von Idealbedingungen ausgehen, etwa multivariat normalverteilten Daten, wie sie in der Forschungspraxis nur selten vorzufinden sind (Cheung & Rensvold, 2002). Zwar betrifft das Problem fehlender Multinormalverteilung auch konfirmatorische Fakto-

renanalysen als solche und hat dort zur Entwicklung robuster Schätzer geführt, wie sie auch im Rahmen dieser Studie Anwendung finden (vgl. Schermelleh-Engel et al., 2003) – allerdings kann ein weitergehender Einfluss fehlender Multinormalverteilung auf *Measurement-Invariance*-Analysen deswegen noch nicht ausgeschlossen werden. Meredith (1993, S. 540) kommt aufgrund solcher Limitierungen zu dem Fazit:

> It should be obvious that measurement invariance[s] … are idealizations. They are, however, enormously useful idealizations in their application to psychological theory building and evaluation. Their validity and existence in the real world of psychological measurement and research can never be finally established in practice.

Gerade für Analysen einzelner Fragebogenitems liefern darüber hinaus spezialisierte Methoden aus dem varianzanalytischen Bereich genauere Ergebnisse (Byrne & Watkins, 2016), da die Analyse der *Measurement Invariance* immer nur Grenzwerte für das gesamte Modell ausgibt: Dies kann dazu führen, dass kleinere Abweichungen einzelner Items unentdeckt bleiben, solange sie isoliert auftreten und nur einzelne Fragen betreffen. Dennoch gilt das *Measurement-Invariance*-Verfahren jenseits derartiger Einzelitemanalysen weitgehend als der „most powerful and versatile approach for testing measurement invariance“ (Milfont & Fischer, 2010).

Die Stärke der *Measurement Invariance* liegt nicht nur darin, dass sie die Kriterien für die Annahme eines Messmodells um eine länderspezifische Komponente ergänzt. Sie ermöglicht es zugleich, Fragen der konzeptuellen und der operationalen Äquivalenz innerhalb der statistischen Methode selbst zu verhandeln (Cheung & Rensvold, 2002). Somit stellt eine *Measurement-Invariance*-Analyse Messäquivalenz als elementares, in die Hauptuntersuchung inkorporiertes Analyseelement von Anfang an bereit: Die mehrstufige Validierung erlaubt zwar keine eindeutige Antwort auf die hermeneutische Frage, was das gemessene Gemeinsame ist – aber sie erlaubt es, mit hoher Wahrscheinlichkeit davon auszugehen, dass es ein Gemeinsames gibt, über das dort verhandelt wird. Sie ermöglicht es, gleichzeitig zu überprüfen, ob beobachtete Differenzen an den Gruppen selbst oder am verwendeten Instrument liegen (Brown, 2015), etwa weil sich Bedeutungszuschreibungen zwischen Ländern ändern, und kann, unter Berücksichtigung der genannten Limitierungen, eine verhältnismäßig breite Palette verschiedener Bias abdecken. Sie zeigt nicht nur, wo sich gemeinsame Skalen bilden lassen, sondern auch, wo Unterschiede zu groß sind, um von einem gemeinsamen Instrument erfasst zu werden. Gerade in den Fällen, in denen sich eben keine gemeinsamen Skalen bilden lassen, wo die strukturellen Unterschiede zwischen den einzelnen Untersuchungsländern oder kulturspezifisch unterschiedliche Bedeutungszuweisungen zu grundsätzlich werden, um sie statistisch abzusichern, zeigt die Methode entsprechende Leerstellen auf, die quantitativ nicht mehr aufzufüllen sind (Cheung & Rensvold, 2002).

Dies sollte jedoch nicht unbedingt als Schwäche der Methodik gewertet werden, vielmehr zeigt es, dass sie nicht nur in der Lage ist, Unterschiede zwischen Ländern valide abzubilden, sondern ebenso scharf zu erkennen, wo die strukturellen Unterschiede, etwa zwischen den Studienbedingungen, zu groß, zu elementar werden, um mittels des Fragebogens erfasst zu werden:

> Metric invariance, for example, need not be seen merely as an obstacle that must be surmounted before the equality of latent means can be assessed; rather, it should be seen as a source of potentially interesting and valuable information about how different groups view the world (Cheung & Rensvold, 2002, S. 252).

Hier ermöglicht es ein Paradigma, einen Blick auf die Grenzen seiner eigenen Analysefähigkeit zu werfen und mit dem Vergleich selbst zugleich auch die Grenzen seiner Einsatzfähigkeit zu reflektieren. Bereits die Feststellung dieser Grenzen, etwa ob sich ein Konstrukt ausschließlich in OECD-Ländern validieren lässt oder ob es auf weitere Kontexte generalisierbar ist, kann eine wesentliche wissenschaftliche Erkenntnis darstellen (vgl. Lange, 2015). Gerade im Fall wesentlicher konzeptueller Unterschiede kann es sich als inhaltlich wenig sinnvoll erweisen, eine Messung mittels gemeinsamer Skalen vornehmen zu wollen, sodass alternative Forschungsmethoden in solchen Fällen eine bessere Herangehensweise darstellen. Dementsprechend ist die hier festgestellte Grenze messtheoretischer, nicht epistemologischer Natur, ist Mess-, nicht Forschungsgrenze. Das gilt insbesondere, da die Grenzen der Vergleichbarkeit mittels deduktiv-quantitativer Methoden nicht die Grenzen der Vergleichbarkeit insgesamt darstellen. Die Alterität von Konstrukten, die im Rahmen der hier vorgestellten Methodik lediglich als Messfehler sichtbar wird, kann selbstverständlich selbst wiederum eine wesentliche Kategorie ländervergleichender Forschung sein und sollte Anstoß zu weiteren, methodisch anders angelegten Untersuchungen geben. Quantitativ-explorative und qualitative Methoden können weitergehende Vergleiche auch in Bereichen ermöglichen, die mittels deduktiv-quantitativer Methodik aufgrund ihrer Diversität nicht analysierbar sind.

Zugleich kann die *Measurement-Invariance*-Messung, indem sie nicht nur Unterschiede betont, sondern auch statistische Hinweise auf Gemeinsames, Vergleichbares liefert und die grundsätzliche Gleichwertigkeit aller untersuchten Systeme voraussetzt, einen Beitrag zur Reflexion der Nostri- bzw. Altrifizierung durch den Forschenden leisten. Sie kann helfen, während der Skalenkonstruktion den eigenen *Cultural Bias* zu überprüfen, ethnozentrische Überinterpretationen zu vermeiden und das Ausmaß an Alterität bei der Übertragung von quantitativen Forschungsvorhaben von einem räumlichen Kontext in einen anderen zu reflektieren. Gerade solche Reflexionen jedoch sind notwendig, um die Validität von Vergleichen sicherzustellen.

Literatur

Alm, F., Jungert, T. & Thornberg, R. (2014). *Nyantagna lärarstudenters motiv, motivation, självtillit och akademiska engagemang* [Die Motive, Motivation, Selbstwirksamkeit und das akademische Engagement beginnender Lehramtsstudierender]. Linköping: Linköpings Universitet.

Backhaus, K., Erichson, B. & Weiber, R. (2015). *Fortgeschrittene Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (3., überarbeitete und aktualisierte Aufl.). Berlin: Gabler.

Bartlett, L. (2016). *Rethinking case study research. A comparative approach*. Florence: Taylor and Francis.

Bastick, T. (2000). *The measurement of teacher motivation: Cross-cultural and gender comparesons*. Paper presented at the Annual Meeting of the Society for Cross-Cultural Research, 29th February, New Orleans, LA.

Beaujean, A.A. (2014). *Latent variable modeling using R. A step by step guide*. New York: Routledge.

Berry, J.W., Poortinga, Y.H., Segall, M.H. & Dasen, P.R. (2002). *Cross-cultural psychology. Research and applications* (2nd ed.). Cambridge, UK: Cambridge University Press.

Blömeke, S. (2011). Überzeugungen in der Lehrerausbildungsforschung. Wie lässt sich dasselbe in unterschiedlichen Kulturkreisen messen? *Beiträge zur Lehrerinnen- und Lehrerbildung, 29* (1), 53–65.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Byrne, B.M. & Watkins, D. (2016). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34* (2), 155–175.

Cheung, G.W. & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling, 9* (2), 233–255.

Dale, R. & Robertson, S. (2009). Beyond methodological 'ISMS' in comparative education in an era of globalisation. In R. Cowen & A.M. Kazamias (Eds.), *International handbook of comparative education* (pp. 1113–1127). Dordrecht: Springer Netherlands.

European Commission. (2017). *Eurydice*. Verfügbar unter https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Main_Page [28.07.2017].

European Commission, EACEA & Eurydice. (2016). *Teachers' and school heads' salaries and allowances in Europe – 2015/16. Eurydice facts and figures*. Luxembourg: Publications Office of the European Union.

Eurydice. (2017a). *Germany. Teachers and education staff*. Verfügbar unter https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Germany:Teachers_and_Education_Staff [09.05.2017].

Eurydice. (2017b). *Sweden. Teachers and education staff*. Verfügbar unter https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Sweden:Teachers_and_Education_Staff [09.05.2017].

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Games, P.A. & Howell, J.F. (1976). Pair wise multiple comparison procedures with unequal n's and/or variances. *Journal of Educational Statistics, 1*, 13–125.

Gutmann, M. & Rathgeber, B. (2011). Vergleichen und Vergleich in den Wissenschaften. Exemplarische Rekonstruktionen zu einer grundlegenden Handlungsform. In A. Mauz (Hrsg.),

Hermeneutik des Vergleichs. Strukturen, Anwendungen und Grenzen komparativer Verfahren (S. 49–74). Würzburg: Königshausen & Neumann.

Helfrich, H. (2013). *Kulturvergleichende Psychologie*. Wiesbaden: Springer.

Holland, J.L. (1997). *Making vocational choices. A theory of vocational personalities and work environments*. Odessa, FL: Psychological Assessment Resources.

Hu, L.-T. & 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.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34* (2), 183–202.

Kaub, K., Stoll, G., Biermann, A., Spinath, F.M. & Brünken, R. (2014). Interessenkongruenz, Belastungserleben und motivationale Orientierung bei Einsteigern im Lehramtsstudium. *Zeitschrift für Arbeits- und Organisationspsychologie A&O, 58* (3), 125–139.

Keller-Schneider, M. (2011). Die Bedeutung von Berufswahlmotiven von Lehrpersonen in der Bewältigung beruflicher Anforderungen in der Berufseingangsphase. *Lehrerbildung auf dem Prüfstand, 4* (2), 157–185.

Keller-Schneider, M., Weiß, S. & Kiel, E. (im Druck). Warum Lehrer/in werden? Idealismus, Sicherheit oder ‚da wusste ich nichts besseres'? Ein Vergleich von Berufswahlmotiven zwischen deutschen und schweizerischen Lehramtsstudierenden und die Bedeutung von länderspezifischen Bedingungen. *Schweizerische Zeitschrift für Bildungswissenschaften*.

König, J., Rothland, M., Darge, K., Lünnemann, M. & Tachtsoglou, S. (2013). Erfassung und Struktur berufswahlrelevanter Faktoren für die Lehrerausbildung und den Lehrerberuf in Deutschland, Österreich und der Schweiz. *Zeitschrift für Erziehungswissenschaft, 16* (3), 553–577.

Kyriacou, C., Kunc, R., Stephens, P. & Hultgren, Å. (2003). Student teachers' expectations of teaching as a career in England and Norway. *Educational Review, 55* (3), 255–263.

Lange, S. (2015). Methodische Reflexionen zur Teilhabe von Ländern der Entwicklungszusammenarbeit an internationalen Vergleichsstudien. *Zeitschrift für internationale Bildungsforschung und Entwicklungspädagogik, 38* (4), 16–24.

Larcher, S. & Oelkers, J. (2004). Deutsche Lehrerbildung im internationalen Vergleich. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Hrsg.), *Handbuch Lehrerbildung* (S. 128–150). Bad Heilbrunn: Klinkhardt.

LMU München. (o.J.). *Lehramt Gymnasium*. Verfügbar unter https://www.uni-muenchen.de/studium/studienangebot/studiengaenge/faecherkombi_lehramt/lehramt_gymnasium/index.html [12.07.2017].

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58* (4), 525–543.

Milfont, T.L. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3* (1), 111–121.

Pädagogische Hochschule Zürich. (o.J.). *Gliederung des Vollzeitstudiums Primarstufe*. Verfügbar unter https://phzh.ch/de/Ausbildung/Studiengaenge/Primarstufe/Bachelorstudiengang-Vollzeit-Primarstufe/Gliederung-des-Studiums/ [25.09.2017].

Parreira do Amaral, M. (2015). Methodologie und Methode der International Vergleichenden Erziehungswissenschaft. In M. Parreira do Amaral & S.K. Amos (Hrsg.), *Internationale und Vergleichende Erziehungswissenschaft* (S. 107–131). Münster: Waxmann.

Pereyra, M.A., Kotthoff, H.-G. & Cowen, R. (2012). *PISA under examination*. Dordrecht: Springer.

Popper, K.R. (1935). *Logik der Forschung*. Wien: Springer.

R Core Team. (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Verfügbar unter https://www.R-project.org/ [27.09.2016].

Rauin, U. & Römer, J. (2010). Motive als Prädiktor des Studien- und Berufserfolgs bei Lehramtsstudierenden und Hauptfachpädagogen. In B. Schwarz, P. Nenniger & R.S. Jäger (Hrsg.), *Erziehungswissenschaftliche Forschung – nachhaltige Bildung* (S. 244–252). Landau: Verlag Empirische Pädagogik.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. Version 0.5-21. *Journal of Statistical Software, 48* (2), 1–36.

Rothland, M. (2014). Warum entscheiden sich Studierende für den Lehrerberuf? In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (2., überarbeitete und erweiterte Aufl.) (S. 349–385). Münster: Waxmann.

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.

Schmitz, E. & Leidl, J. (1999). Brennt wirklich aus, wer entflammt war? Studie 2: Eine LISREL-Analyse zum Burnout-Prozess bei Lehrpersonen. *Psychologie in Erziehung und Unterricht, 46* (4), 302–310.

Schulz-Heidorf, K. & Solheim, O.J. (2016). Adapted teaching: A chance to reduce the effect of social origin? A comparison between Germany and Norway, using PIRLS 2011. *Tertium Comparationis, 22* (2), 230–260.

semTools Contributors. (2016). *semTools: Useful tools for structural equation modeling*. Verfügbar unter http://cran.r-project.org/package=semTools [29.09.2016].

Södertörn Högskola [Universität Södertörn] (o.J.). *Ämneslärarutbildning med interkulturell profil med inriktning mot gymnasieskolan* [Fachlehrerausbildung mit interkulturellem Profil mit dem Schwerpunkt Sekundarschulen]. Verfügbar unter http://www.sh.se/p3/ext/content.nsf/aget?openagent&key=sh_program_page_0_P4208 [12.07.2017].

Steiner-Khamsi, G. (2002). Re-framing educational borrowing as a policy strategy. In M. Caruso & H.-E. Tenorth (Hrsg.), *Internationalisierung. Semantik und Bildungssystem in vergleichender Perspektive* (S. 57–89). Frankfurt a.M.: Lang.

Stiegelbauer, S. (1992). *Why we want to be teachers: New teachers talk about their reasons for entering the profession*. Paper presented at the Annual Meeting of the American Educational Research Association, April 20–24, San Francisco. Verfügbar unter http://files.eric.ed.gov/fulltext/ED348367.pdf [25.09.2017].

Suryani, A., Watt, H.M.G. & Richardson, P.W. (2016). Students' motivations to become teachers. FIT-Choice findings from Indonesia. *International Journal of Quantitative Research in Education, 3* (3), 179–203.

Syring, M., Weiß, S., Keller-Schneider, M., Hellsten, M. & Kiel, E. (2017). Berufsfeld ‚Kindheitspädagoge/in': Berufsbilder, Professionalisierungswege und Studienwahlmotive im europäischen Vergleich. *Zeitschrift für Pädagogik, 63* (2), 139–162.

Ulich, K. (2004). *Ich will Lehrer/in werden. Eine Untersuchung zu den Berufsmotiven von Studierenden*. Weinheim: Beltz.

Watt, H. & Richardson, P. (2007). Motivational factors influencing teaching as a career choice: Development and validation of the 'FIT-Choice' scale. *Journal of Experimental Education, 75* (3), 167–202.

Weiß, S. & Kiel, E. (2013). Who chooses primary teaching and why. *Issues in Educational Research, 23* (3), 415–433.

Zumwalt, K. & Craig, E. (2008). Who is teaching? Does it matter? In M. Cochran-Smith, S. Feiman-Nemser, D.J. McIntyre & K.E. Demers (Eds.), *Handbook of research on teacher education. Enduring questions and changing contexts* (3rd ed.) (pp. 404–423). New York: Routledge.