Auf dem Weg zum „Hochschul-PISA“?

Zur Messung „soziologischer Kompetenzen“

Von Felix Wolter und Jürgen Schiener

Zusammenfassung: Im Gegensatz zum allgemeinbildenden schulischen Bereich steckt die Kompetenzdiagnostik im Hochschulbereich immer noch in den Kinderschuhen. Dies gilt auch für die Soziologie, wobei dies verwunderlich ist, da gerade in dieser Disziplin die Methodenkompetenz und das Bewusstsein für die Notwendigkeit valider Indikatoren vergleichsweise hoch sein dürften. Auch die Entwicklungen rund um den Bologna-Prozess und die aktuelle Diskussion über das sog. CHE-Ranking legen nahe, dass ein erheblicher Bedarf an validen Messinstrumenten zum tatsächlichen Kompetenzerwerb von Studierenden besteht. Im vorliegenden Beitrag werden daher die Möglichkeiten einer Definition und standardisierten Messung „soziologischer Kompetenzen“ ausgelotet. Neben der evaluatorischen Motivation für eine Kompetenzmessung im Hochschulbereich ist diese auch grundsätzlich für soziologische Fragestellungen von einigem Interesse, wenn es beispielsweise um Determinanten und Erträge des Bildungserwerbs *jenseits* von Noten oder Zertifikaten geht. Präsentiert werden neben grundsätzlichen Überlegungen Ergebnisse und Erkenntnisse einer Pilotstudie, in deren Rahmen 540 Soziologiestudierende einer Kompetenzmessung im Stil der PISA-Studien unterzogen wurden. Soziologische Kompetenzen sind dabei als fachspezifische kognitive Leistungsdispositionen konzipiert, dimensional untergliedert, mit entsprechenden Testaufgaben operationalisiert und mit Verfahren der Item-Response-Theorie skaliert. Die Befunde demonstrieren nicht nur die grundsätzliche Durchführbarkeit vergleichbarer Vorhaben, sondern weisen auch auf gute Skaleneigenschaften des hier vorgestellten Kompetenzindikators hin. In ausgewählten inhaltlichen Analysen zeigt sich zudem eine hohe externe Validität des Indikators. Die Ergebnisse sprechen dafür, die Idee einer standardisierten Kompetenzmessung in der Soziologie in Zukunft weiterzuverfolgen und auszubauen.

1. Einleitung[1]

Seit einigen Jahren ist das Thema Kompetenzdiagnostik – also die Definition, Modellierung, Messung und Evaluation von fachlichen Kompetenzen im schulischen, beruflichen und universitären Bereich – ein in den Sozialwissenschaften intensiv beforschtes Gebiet (für viele: Prenzel et al. 2008). Kompetenzen werden dabei meist als (kognitive) Handlungsdispositionen begriffen, die Akteure in die Lage versetzen, angemessen auf fachspezifische Anforderungen oder Problemsituationen zu reagieren. In mehreren großen internationalen und nationalen Erhebungen werden mittlerweile kontinuierlich Kompetenzen von Schülern erhoben – zu nennen sind neben den PISA-Studien (Baumert et al. 2001; Prenzel et al. 2008) beispielsweise die TIMSS- (Baumert et al. 1998) und IGLU-Erhebungen (Bos et al. 2007) sowie der IQB-Ländervergleich (Stanat et al. 2012).[2] Zur Messung von Kompetenzen im allgemeinbildenden schulischen Bildungssystem gibt es mittlerweile gut ausgearbeitete Kompetenzmodelle und Messinstrumente.

1 Die vorliegende Studie wurde im Rahmen eines zweisemestrigen Lehrforschungsprojekts am Institut für Soziologie der Universität Mainz durchgeführt. Wir danken den Teilnehmerinnen und Teilnehmern des Projektseminars für ihren Einsatz und drei anonymen Gutachtern sowie Peter Preisendörfer und Ingmar Ehler für wertvolle Hinweise zum Manuskript.

2 PISA: „Programme for International Student Assessment“; TIMSS: „Trends in International Mathematics and Science Study”; IGLU: „Internationale Grundschul-Lese-Untersuchung”; IQB: „Institut zur Qualitätsentwicklung im Bildungswesen”.

Anders stellt sich hingegen die Situation dar, wenn es um Kompetenzen im Hochschulbereich geht. In vielen Fachgebieten, und auch in der Soziologie, steht die Entwicklung von Kompetenzmodellen und Messinstrumenten zur Erfassung der fach- bzw. domänenspezifischen Kompetenzen, die in universitären Studiengängen vermittelt werden, noch aus (Blömeke 2013; Förster et al. 2012; Zlatkin-Troitschanskaia et al. 2012). Dass aber ein erheblicher Bedarf für eine valide Kompetenzmessung besteht, wird nicht zuletzt durch die Diskussion um Hochschulrankings und Evaluationen im Hochschulbereich illustriert (DGS 2012). Rankings wie z.B. auch das problematische CHE-Ranking sind entweder *input-orientiert*, also gerichtet auf Strukturen und Prozesse von Studium und Lehre, wie z.B. die sachliche und personale Ausstattung von Instituten, oder sie basieren auf subjektiven Indikatoren wie Einschätzungen von Studierenden zu ihrem Lernerfolg, zu Dozenten oder allgemein zur Qualität der Lehre (vgl. z.B. Braun et al. 2008). Unseres Erachtens sinnvoller wäre aber eine *Output-Orientierung*, die in einer Art „Hochschul-PISA" anhand valider, *objektiver* Indikatoren den tatsächlichen Bildungserfolg der universitären Einrichtungen, und zwar konkret den Kompetenzgewinn der Studierenden, betrachtet (Zlatkin-Troitschanskaia et al. 2012). Schließlich ist die Kompetenzorientierung der neuen BA- und MA-Studiengänge eine wesentliche Zielgröße des Bologna-Prozesses und die Curricula zielen auf den Erwerb bestimmter, in Modulhandbüchern und Veranstaltungsbeschreibungen festgelegter Kompetenzen ab, deren Abschätzung anhand von systematisch geprüften Messinstrumenten in den meisten Fällen aber (noch) nicht möglich ist.

Neben dieser evaluatorischen Motivation für eine valide Kompetenzmessung gibt es aber auch eine inhaltliche bzw. soziologische Forderung nach validen Maßen für Bildungserfolg, wenn es beispielsweise um Fragen nach Determinanten (z.B. soziale Herkunft) und Erträgen (z.B. Lohneinkommen) von Bildung *jenseits* von Bildungszertifikaten oder Zensuren geht. Konventionelle Leistungsmaße wie Noten oder Bildungszertifikate, die in der einschlägigen Forschung in der Regel herangezogen werden, sind u.a. deshalb nur bedingt geeignete (Proxy-)Indikatoren, weil sie einem erheblichen Spielraum der vergebenden Akteure unterliegen, nach institutionellen Charakteristika des jeweiligen Bildungs(sub)systems variieren und daher letztlich nicht objektiv (oder zumindest intersubjektiv) und vergleichbar sind (z.B. Wissenschaftsrat 2012). Für den Hochschulbereich haben Müller-Benedict und Tsarouha (2011) diese Probleme besonders instruktiv nachgewiesen.

Und es gibt noch eine dritte Motivation, eine Debatte über soziologische Kompetenzen anzustoßen: Das Berufsbild eines Soziologen bzw. einer Soziologin, die Inhalte der soziologischen Ausbildung und eben die dort erworbenen Kompetenzen sind in großen Teilen der Öffentlichkeit und der Berufswelt oft immer noch vage und unklar. Daher wäre es auch für die Außendarstellung der Soziologie sinnvoll und dienlich, wenn es gelänge, das Wissen und die Fähigkeiten einer Soziologin / eines Soziologen sprachlich zu fixieren und deren Messbarkeit zu demonstrieren.

Während Projekte, die sich der Messung von Kompetenzen im Hochschulbereich widmen,[3] in einigen Fachgebieten bereits erste empirische Ergebnisse vorweisen können, ist das für die Soziologie bisher nicht der Fall. Dies ist verwunderlich, da zum einen die Methodenkompetenz und die Erfahrung mit Messproblemen, und damit auch das Bewusstsein für den Bedarf an validen Indikatoren, in dieser Disziplin vergleichsweise hoch sein dürften. Zum anderen läuft aktuell in diesem Fach die bereits erwähnte Diskussion um Hochschulevaluation, verbunden mit einer teils harschen Kritik an bestehenden Indikatoren, wie sie beispielsweise das CHE-Ranking heranzieht (DGS 2012). Allerdings hat diese Kritik bisher nicht zu konkreten Vorschlägen geführt, die bessere Alternativen bieten.

3 Vgl. Abschnitt 2.3 für eine Darstellung des Forschungsstands.

Daher möchten wir in diesem Beitrag die Frage nach einer Kompetenzmessung in der Soziologie aufwerfen und erste Erkenntnisse und Ergebnisse aus einer Pilotstudie präsentieren. Grundsätzlich umfasst die Problemstellung zwei Aufgaben: Erstens bedarf es einer theoretisch-konzeptionellen Diskussion, ob und wie sich „soziologische Kompetenzen" definieren, abgrenzen und in einem Kompetenzmodell festschreiben lassen. Zweitens ist das Kompetenzmodell einer empirischen Messung zugänglich zu machen, wobei sich hier im Bereich der Kompetenzdiagnostik die *Item-Response-Theorie (IRT)* und insbesondere das Rasch-Modell als zentrale methodische Herangehensweise etabliert hat.

Bei diesem Vorhaben ist von vornherein klar, dass es angesichts der multiparadigmatischen Ausrichtung der Disziplin, der vielen nebeneinander her arbeitenden speziellen Soziologien und der Vorbehalte, die in einigen Bereichen des Faches gegenüber standardisierten quantitativen Messungen bestehen, kein leichtes sein wird. Dennoch plädieren wir dafür, sich der Aufgabe pragmatisch und unvoreingenommen zu stellen. Nicht zuletzt besteht ja ein Konsens unter allen Fachvertretern, dass ein Bedarf an Leistungsmessungen vorhanden ist, und auch die grundsätzliche Durchführbarkeit wird von allen akzeptiert – denn an jedem deutschen soziologischen Institut und von vermutlich jedem Lehrenden in der Soziologie werden Prüfungen in Form von Klausuren, Hausarbeiten oder mündlichen Prüfungen durchgeführt und bewertet, also letztlich Kompetenzmessungen angestrebt, obwohl die psychometrischen Eigenschaften der eingesetzten Instrumente in den seltensten Fällen untersucht werden.

Im Folgenden gehen wir zunächst auf den Forschungsstand im Bereich der Kompetenzdiagnostik ein und geben einen Überblick über Begriffe und Konzepte, über für Kompetenzmessungen zentrale Methoden der IRT sowie über empirische Forschungsliteratur zur Kompetenzmessung im Hochschulbereich (Abschnitt 2). Im darauffolgenden Abschnitt berichten wir von einer an der Universität Mainz durchgeführten Pilotstudie zur Messung soziologischer Kompetenzen. Zunächst werden Überlegungen zu einem Kompetenzmodell sowie das Design und der Ablauf der Erhebung präsentiert (Abschnitt 3). Der sich anschließende Abschnitt stellt zuerst empirische Ergebnisse zur Skalierung unserer Kompetenzindikatoren vor; sodann wird anhand einiger ausgewählter inhaltlicher Analysen das Analysepotenzial und die externe Validität (Konstruktvalidität) eines globalen Kompetenzindikators aufgezeigt (Abschnitt 4). Unser Beitrag schließt mit einer Zusammenfassung der wichtigsten Erkenntnisse aus der Pilotstudie, Schlussfolgerungen und einem Ausblick für die künftige Forschung.

2. Kompetenzdiagnostik: Begriffe, Methoden und Forschungsstand

Die sehr umfangreiche Literatur zu Kompetenzdiagnostik lässt sich in drei Teilbereiche einteilen: In einem ersten Forschungsstrang gehen theoretisch-konzeptionelle Arbeiten Fragen nach geeigneten Kompetenzdefinitionen bzw. -konzepten nach (z.B. Baumert / Kunter 2006; Klieme / Hartig 2007; Weinert 2001). Hier geht es auch um die theoretische Anschlussfähigkeit und Integrierbarkeit von Kompetenzkonstrukten hinsichtlich bestehender pädagogischer, (kognitions-)psychologischer und soziologischer Theorien (z.B. Frosch 2012; Zlatkin-Troitschanskaia / Seidel 2011). Ein zweiter Teilbereich umfasst methodisch-statistische Arbeiten. Forschungsgegenstand sind hier statistische Modelle – meist IRT-Verfahren – zur empirischen Skalierung von Kompetenzen (z.B. Geiser / Eid 2010; Moosbrugger 2012; Rost 2004; Wilson 2005). In den dritten Teilbereich lassen sich schließlich empirische Umsetzungen bzw. Anwendungen von Kompetenzmessungen einordnen – Arbeiten, die entweder neue Skalen entwickeln und empirisch prüfen (z.B. Förster / Zlatkin-Troitschanskaia 2010; Winkelmann / Robitzsch 2009), oder solche, die bereits etablierte Skalen zur empirischen Anwendung bringen (vgl. die oben zitierten Quellen zu PISA und anderen Erhebungen). Im Folgenden gehen wir kurz auf den Stand der Literatur zu den drei Teilbereichen ein.

2.1 Zum Begriff der Kompetenz

Weil die Kompetenzdiagnostik interdisziplinär in der Pädagogik, Psychologie, Soziologie, den Wirtschaftswissenschaften und auch im kulturwissenschaftlich-philosophischen Feld betrieben wird, existiert eine Vielzahl an Herangehensweisen und Definitionsangeboten zum Begriff der Kompetenz (siehe Klieme / Hartig 2007 für eine ausführliche Darstellung). Da hier keine erschöpfende Diskussion erfolgen kann, empfiehlt es sich, pragmatisch vorzugehen. Eine Definition, die empirisch anschlussfähig ist und als Grundlage für viele Studien (z.B. Förster / Zlatkin-Troitschanskaia 2010) fungiert, versteht Kompetenzen als „*kontextspezifische kognitive Leistungsdispositionen*, die sich funktional auf Situationen und Anforderungen in bestimmten *Domänen* beziehen" (Klieme / Leutner 2006: 879, Herv. im Orig.; vgl. auch Weinert 2001). Nach dieser Definition ist Kompetenz die Fähigkeit bzw. das Wissen und Können, in einem genau abgegrenzten fachlichen Bereich (Domäne), in einem bestimmten Kontext – etwa im Rahmen eines Hochschulstudiums oder einer bestimmten Lehrveranstaltung – auf eine konkrete Situation – z.B. eine Problemstellung oder Aufgabe – angemessen reagieren zu können. Kompetenzen in diesem Sinne sind also *Handlungsfähigkeiten*, „Can-Do-Aussagen" (Pant et al. 2012: 50), die sich auf einen spezifischen Kontext beziehen bzw. deren Binnenstruktur sich aus konkreten situationellen Anforderungen ergibt (Hartig / Klieme 2006: 130 f). Zudem sind sie grundsätzlich erlernbar, wodurch sie sich zu Konzepten, die allgemeine, nicht erlernbare Dispositionen beschreiben (wie z.B. Intelligenz), abgrenzen (Klieme / Leutner 2006: 879). Viele Autoren (z.B. Weinert 2001) beziehen neben dieser Fähigkeits- oder kognitiven Komponente auch eine motivational-volitionale Komponente in den Kompetenzbegriff mit ein, die eine positiv-aufgeschlossene Einstellung zur fachlichen Domäne und den entsprechenden Problemsituationen beinhaltet. Ein diesbezüglich häufig herangezogener Kompetenzbegriff geht auf Roth (1971) zurück, für den Kompetenz erstens *Sachkompetenz* ist, welche sich wiederum in Fach- und Methodenkompetenz untergliedert und das Wissen und Können in einer bestimmten Domäne beschreibt, zweitens eine *Selbstkompetenz* beinhaltet, welche Selbstregulationsfähigkeiten wie Lernstrategien, Selbstwirksamkeitserwartung, Leistungsbereitschaft oder Evaluationskompetenz (Klieme / Hartig 2007: 20) beschreibt, und drittens auch eine *Sozialkompetenz* umfasst, die sich auf Fähigkeiten wie Verantwortungsbewusstsein in Gruppen oder den Willen bezieht, die Sachkompetenzen sozial zu artikulieren und „Verantwortlichkeit" zu beanspruchen. Für empirische Umsetzungen empfehlen Klieme / Leutner (2006: 880) allerdings, die kognitive Komponente (also Fach- und Methodenkompetenz) getrennt von den anderen beiden Komponenten zu erfassen.

Hartig und Klieme (2006) unterscheiden weiterhin zwischen *Kompetenzstrukturmodellen* und *Kompetenzniveaumodellen* (bzw. begrifflich gleichbedeutend *Kompetenzstufenmodellen*). Erstere beziehen sich auf die inhaltliche Dimensionierung des konkreten Kompetenzbegriffs, zweitere auf eine verbale Objektivierung unterschiedlicher (unterschiedlich hoher) Ausprägungen der jeweiligen Kompetenz. Die Festschreibung eines Kompetenzstrukturmodells orientiert sich an der jeweiligen fachlichen Domäne – beispielsweise untergliedert der IQB-Ländervergleich (Stanat et al. 2012) Kompetenzen von Viertklässlern im Fach Deutsch in die Bereiche „Sprechen und Zuhören", „Schreiben", „Lesen – mit Texten und Medien umgehen" sowie „Sprache und Sprachgebrauch untersuchen" (Böhme et al. 2012). Kompetenzniveaumodelle werden in der Regel ex post, also nach der Skalenentwicklung, auf empirischer Grundlage entwickelt, indem die Skala in Bereiche unterteilt wird und diese sodann sprachlich mit Begriffen belegt werden (Hartig / Klieme 2006: 133 f). Aus Platzgründen werden wir die Frage nach einer Definition von Kompetenzniveaus im vorliegenden Beitrag nicht weiterverfolgen.

Eine Möglichkeit, eine Kompetenzstruktur zu konzeptionalisieren und hinsichtlich ihrer fachlichen Gegenstände und Fähigkeiten zu hierarchisieren, ist die „Bloom'sche Taxonomie"

(Bloom 1956) bzw. deren jüngere Revision (Anderson / Krathwohl 2001; Krathwohl 2002). In der ersten Version der Taxonomie (die folgende Darstellung orientiert sich an Krathwohl 2002) wurden die kognitiven Fähigkeiten *Wissen*, *Verstehen*, *Anwenden*, *Analysieren*, *Synthetisieren* und *Evaluieren* abgegrenzt. Entscheidend ist die Annahme, dass die Stufen hierarchisch angeordnet sind und das Beherrschen einer höheren Stufe das Beherrschen aller niedrigeren Stufen erfordert. In der Revision wurde die Taxonomie erweitert und erstreckt sich nun entlang zweier Dimensionen, wobei sich die erste Dimension auf (Wissens-) Gegenstände und die zweite auf kognitive Prozesse bezieht. Bei den Wissensgegenständen wird zwischen *Faktenwissen*, *konzeptionellem Wissen*, *prozeduralem Wissen* und *metakognitivem Wissen* unterschieden. Die kognitiven Prozesse sind *Erinnern*, *Verstehen*, *Anwenden*, *Analysieren*, *Bewerten* und *Erschaffen*. Die zwei Dimensionen mit ihren jeweiligen Hauptausprägungen können gekreuzt werden, so dass eine Matrix mit 24 Feldern entsteht. Im weiteren Vorgehen wären diese Felder jeweils mit Testitems zu bestücken, welche die jeweilige gegenstandsbezogene Fähigkeit für eine bestimmte Domäne messen.[4]

Obgleich die Bloom-Taxonomie keine empirische Grundlage hat und es in der Praxis schwierig werden dürfte, die 24-Felder-Matrix empirisch umzusetzen bzw., konkret, genügend Items für jede einzelne Zelle zu testen, kann das Konzept u.E. durchaus die Formulierung eines Kompetenzmodells kanalisieren, strukturieren und überhaupt eine Vorstellung davon vermitteln, welche Fähigkeiten in das Kompetenzmodell eingehen könnten. Ein Beispiel für eine empirische Umsetzung der Bloom-Taxonomie ist die TEDS-LT-Studie[5] zur Messung von fachlichen, fachdidaktischen und pädagogisch-psychologischen Kompetenzen von angehenden Lehrern. Hier wurde die revidierte Bloom-Taxonomie in einer stark kondensierten Fassung angewendet und nur zwischen drei Stufen (Erinnern und Abrufen, Verstehen und Anwenden, Bewerten und Generieren von Handlungsoptionen) unterschieden (Blömeke et al. 2011: 15ff).

Auf einen Versuch, vor diesem Hintergrund ein Strukturmodell für „soziologische Kompetenzen" zu entwickeln, gehen wir in Abschnitt 3 ein.

2.2 Item-Response-Theorie

Für die empirische Umsetzung von Kompetenzmessungen haben sich mittlerweile die Methoden der Item-Response-Theorie (IRT) durchgesetzt. Unter dem Begriff wird eine ganze Reihe an Verfahren subsumiert, deren Ziel es ist, aufgrund von beobachtbaren Probandenantworten auf Testitems auf zugrundeliegende und nicht direkt beobachtbare latente Merkmale – etwa eine bestimmte Kompetenz – zu schließen. Die IRT-Verfahren grenzen sich dabei zur klassischen Testtheorie dadurch ab, dass sie wesentlich strengere Anforderungen an die Tests bzw. Messinstrumente stellen (Moosbrugger 2012: 229) und sowohl Aufgabenschwierigkeiten als auch Personenfähigkeiten auf einer gemeinsamen Skala abbilden (Moosbrugger 2012: 233). Das konkrete Ziel der IRT-Verfahren besteht darin, Tests bzw. Testitems zu identifizieren, welche die Anforderung der *Itemhomogenität* bzw. lokalen stochastischen Unabhängigkeit erfüllen: Nur wenn (neben der Schwierigkeit eines Items) ausschließlich die gesuchte latente Eigenschaft bzw. deren Ausprägung für eine bestimmte Versuchsperson die Antworten auf die Testitems bestimmt, misst die resultierende Skala das gewünschte Konstrukt. Ein oft genanntes Beispiel (Strobl 2010: 1) für einen Test, der diese Eigenschaft nicht erfüllt, wäre ein

4 Die Bloom-Hierarchie ist nicht zu verwechseln mit Kompetenzstufen oder -niveaus, da es grundsätzlich möglich ist, auf unterschiedlichen Kompetenzstufen das Spektrum der in der Bloom-Taxonomie beschriebenen Fähigkeiten voneinander zu unterscheiden. So können etwa auch auf einem sehr niedrigen Kompetenzniveau Bewertungs- und Evaluationsfähigkeiten definiert werden, umgekehrt können auch auf einem sehr hohen Kompetenzniveau Fähigkeiten aus dem unteren Bereich der Bloom-Hierarchie (z.B. Wissen) gemessen werden.

5 TEDS-LT: „Teacher Education and Development Study: Learning to Teach".

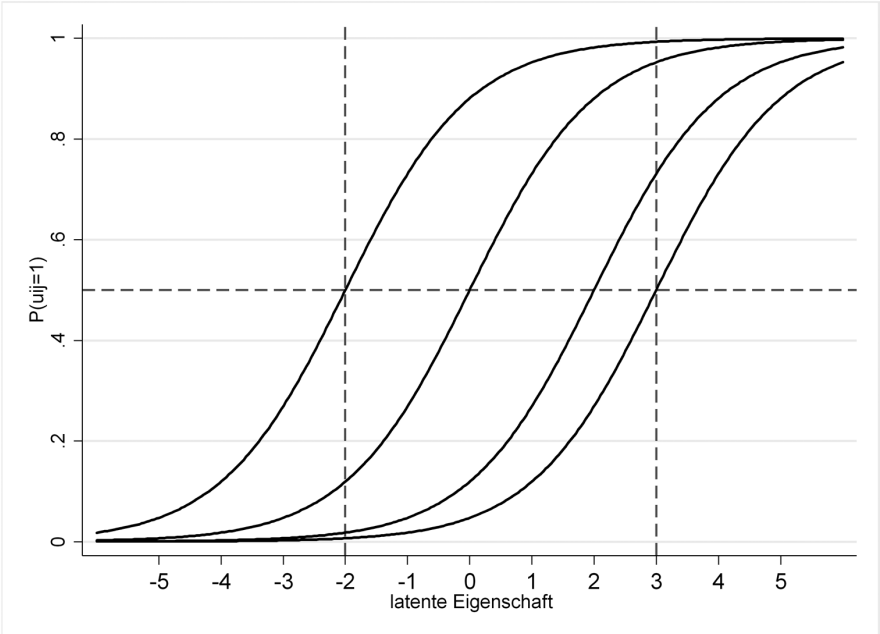
Test auf mathematische Kompetenzen von Schülern, dessen Items jedoch nicht nur die Mathematikkompetenz messen, sondern auch sprachliche Begabung, da einige Aufgaben in Form von Textaufgaben präsentiert werden, die für sprachlich begabtere Schüler leichter zu lösen sind. Anders ausgedrückt überprüfen die IRT-Verfahren die zentrale Annahme der *Eindimensionalität*.

Da hier keine erschöpfende Diskussion aller IRT-Verfahren erfolgen kann, beschränken wir uns darauf, das Rasch-Modell (Rasch 1960) als parametrisches Skalierungsverfahren mit sehr strengen Anforderungen, aber auch den besten Eigenschaften für die resultierende Skala, sowie das verwandte Birnbaum-Modell (Birnbaum 1968) zu erläutern. Eine Rasch-Skala gilt gemeinhin als Optimum einer Skalierung bzw. eines Messinstruments, weshalb dieses Skalierungsverfahren als Standard im Bereich der Kompetenzdiagnostik gesehen werden kann. Im Rasch-Modell wird die Wahrscheinlichkeit einer korrekten Beantwortung eines Items *j* durch eine Testperson *i*, dargestellt durch die Zufallsvariable U_{ij}, durch eine logistische Funktion in Abhängigkeit der Fähigkeit der Testperson θ_i (also der latenten Variable) sowie der Schwierigkeit des Items δ_j (zur Definition von Schwierigkeit siehe unten) beschrieben (Formel 1).

$$P\left(U_{ij}=1\middle|\theta_i,\delta_j\right)=\frac{e^{\theta_i-\delta_j}}{1+e^{\theta_i-\delta_j}} \tag{1}$$

Erfüllen die Daten das Rasch-Modell, so liegt erstens lokale stochastische Unabhängigkeit vor, d.h. die Wahrscheinlichkeit, ein Item korrekt zu lösen, hängt neben einem Zufallsfaktor ausschließlich von dessen Schwierigkeit und der Ausprägung der latenten Eigenschaft ab, zweitens sind die Zeilen- und Spaltenrandsummen der Datenmatrix suffiziente Statistiken für θ_i bzw. für δ_j, was wiederum bedeutet, dass spezifische Objektivität vorliegt und nur die Zahl der gelösten Aufgaben in den Wert der latenten Variablen eingeht und die Auswahl der Aufgaben nebensächlich ist. Die Personenparameter einer Rasch-Skala, also die jeweiligen Ausprägungen der Probanden auf der Fähigkeitsskala, sind intervallskaliert. In Abbildung 1 sind die Itemfunktionen oder *itemcharakteristischen Kurven* (item characteristic curves, ICC) von vier Items dargestellt, die das Rasch-Modell erfüllen. Aus der Darstellung geht auch hervor, dass sich die Items nur hinsichtlich ihrer Schwierigkeit unterscheiden, nicht aber hinsichtlich ihrer Steigung bzw. Trennschärfe. Die Schwierigkeit eines Items wird im Rasch-Modell auf der gleichen Skala wie die Personenparameter gemessen und definiert sich als der Punkt in θ_i, bei dem die Lösungswahrscheinlichkeit für das Item j genau 0,5 beträgt (vgl. die gestrichelten Linien in Abbildung 1).

Abbildung 1: Itemcharakteristische Kurven (ICC) im Rasch-Modell



Quelle: Eigene Darstellung in Anlehnung an Strobl 2010: 11.

Zur Schätzung der Personenparamter θ_i und den Aufgabenparametern δ_j im Rasch-Modell werden Maximum-Likelihood-Verfahren eingesetzt, wobei verschiedene Schätzvarianten möglich sind, deren Darstellung hier nicht erfolgen kann (vgl. hierfür z.B. de Ayala 2009; Strobl 2012 und die dort zitierte Literatur).

Zur Überprüfung, ob ein bestimmter Itemsatz die Bedingungen der Rasch-Skala erfüllt, sind zwei Strategien sinnvoll. Erstens können modellimmanente Tests prüfen, ob sog. *Differential Item Functioning* (DIF) vorliegt, und/oder die Annahme der Eindimensionalität bzw. lokalen stochastischen Unabhängigkeit verletzt ist. Dies geschieht durch Gruppenvergleiche, indem geprüft wird, ob die Schätzung der Aufgabenparameter (Itemschwierigkeiten) in verschiedenen Subgruppen des Datensatzes (signifikant) unter Konstanthaltung der Personenfähigkeit unterschiedlich ausfällt. Ist das der Fall, sind also die Aufgaben für Subgruppen der Daten unterschiedlich schwer zu lösen, so sind die Annahmen des Rasch-Modells nicht erfüllt und problematische Items (oder Personen mit inkonsistenten Antwortmustern) zu entfernen. Modellimmanente itemspezifische Tests werden verwendet, um zu prüfen, ob die beobachteten und die aus dem Modell geschätzten Lösungshäufigkeiten voneinander abweichen.

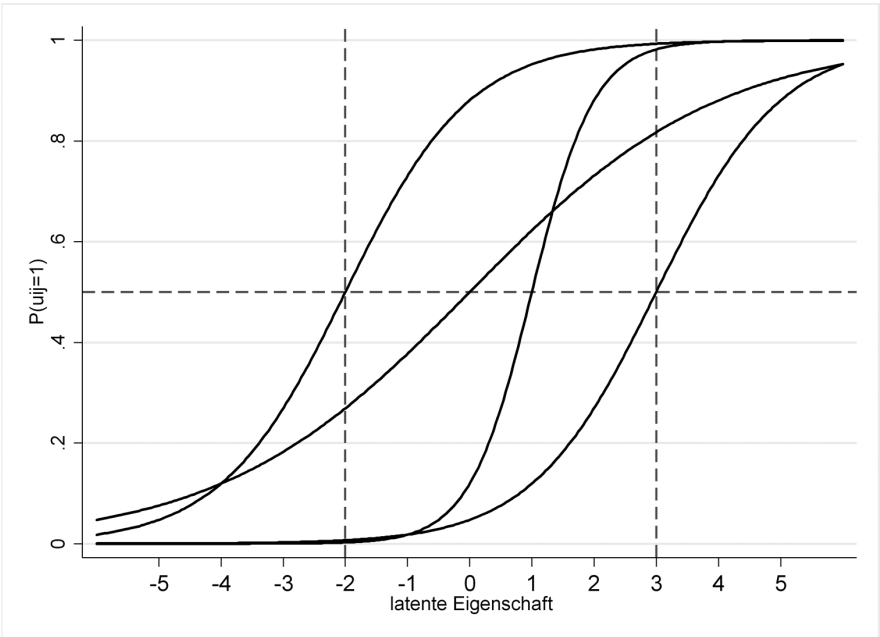
Zur Überprüfung der Modellgeltung kann zweitens die Passung des Raschmodells zu den Daten gegen erweiterte, weniger restriktive Modelle mit laxeren Annahmen getestet werden. Hierfür stehen Likelihood-Ratio-Tests oder Gütemaße wie AIC und BIC zur Verfügung. Stellt sich hierbei heraus, dass alternative Modelle signifikant besser zu den Daten passen, muss das Rasch-Modell streng genommen abgelehnt werden. Allerdings tendieren die Abweichungen verschieden restriktiver Modelle bei hohen Fallzahlen dazu, immer signifikant zu werden (Kubinger / Draxler 2007). Daher beschränken sich beispielsweise auch die PISA-Studien

darauf, nur itemspezifische Teststatistiken zur Überprüfung der Modellgeltung heranzuziehen (vgl. Carstensen 2006: 316).

Eine Erweiterung des Rasch-Modells ist die Aufgabe der Annahme gleicher Trennschärfen bzw. Steigungskoeffizienten der ICC. Erweitert man entsprechend die Modellgleichung um einen zusätzlichen Koeffizienten β_j für jedes Item, erhält man das Birnbaum-Modell oder 2PL-Modell (Formel 2). Wie anhand der in Abbildung 2 dargestellten ICC illustriert wird, können sich diese nunmehr überschneiden, die Reihenfolge der Itemschwierigkeiten verändert sich über θ. Damit gilt für das Birnbaum-Modell keine spezifische Objektivität mehr.

$$P\left(U_{ij}=1\middle|\theta_i,\delta_j,\beta_j\right)=\frac{e^{\beta_j\left(\theta_i-\delta_j\right)}}{1+e^{\beta_j\left(\theta_i-\delta_j\right)}} \tag{2}$$

Abbildung 2: Itemcharakteristische Kurven (ICC) im Birnbaum-Modell



Neben dem Birnbaum-Modell existieren noch andere Erweiterungen und Varianten des Rasch-Modells, die hier allerdings aus Platzgründen nicht diskutiert werden können (vgl. hierfür z.B. de Ayala 2009; Strobl 2010; van der Linden / Hambleton 1997). Hingewiesen sei nur auf das 3PL-Modell, welches zusätzlich zu den variierenden Trennschärfen des Birnbaum-Modells auch einen Rateparameter für jedes Item modelliert, also eine Konstante, die korrekte Lösungen eines Items durch rein zufälliges Beantworten (bzw. Ankreuzen) abbildet.

Strobl (2012: 51) und Ho Yu (2010: 1) verbinden die „Entscheidung" zwischen Rasch- und Birnbaum-Modell (oder erweiterten Modellen) mit unterschiedlichen Forschungszielen: Geht es um die Suche nach einem Modell, das die Daten besonders gut wiedergibt, wäre die Modellierung über ein Birnbaum- oder ein nochmals erweitertes Modell angebracht; geht es hingegen um die Suche nach Daten bzw. Testitems, welche die Anforderungen des Rasch-Modells

erfüllen, sollte das Birnbaum-Modell nicht akzeptiert, sondern sollten Testitems entfernt werden, bis eine Anpassung an das Rasch-Modell erreicht wird. In der Praxis zeigen sich allerdings deutlich unterschiedliche Auffassungen und Denkschulen über die adäquate Modellierungsstrategie. Entgegen dem Primat der Rasch-Befürworter werden beispielsweise die Skalen der TIMSS-Studien grundsätzlich über 3PL-Modelle skaliert (Wu 2010: 21).[6]

2.3 Kompetenzmessung im Hochschulbereich

Was den Forschungsstand zur Kompetenzdiagnostik im Hochschulbereich betrifft, bestimmen aktuell zwei große Initiativen das Forschungsgeschehen. Auf internationaler Ebene wird in der Machbarkeitsstudie AHELO (Assessment of Higher Education Learning Outcomes) der OECD ausgelotet, inwieweit eine länderübergreifende Erhebung von Kompetenzen Studierender im Stil der PISA-Studien realisierbar ist. Hierfür wurden in ersten Pilotstudien 23.000 Studierende aus 17 Ländern (Deutschland ist nicht darunter) befragt, wobei zunächst fächerübergreifende generische Kompetenzen (z.B. Lese- und Problemlösungskompetenzen) und domänenspezifische Kompetenzen in Ökonomie und Ingenieurswissenschaften Gegenstand der Messung waren. Die Ergebnisse sind erst teilweise publiziert (OECD 2013; Tremblay et al. 2012), so dass ein abschließendes Fazit noch aussteht. Bezüglich der Rezeption von AHELO in Deutschland (Braun et al. 2013) besteht unter Kompetenzforschern Einigkeit, dass ein grundsätzlicher Bedarf an Messinstrumenten für eine Kompetenzdiagnostik im Hochschulbereich besteht, gleichzeitig aber „das geeignete Instrument […] noch nicht entwickelt [ist], viele konzeptuelle und definitorische Vorarbeiten […] noch notwendig [sind], bis die Testung selbst vorgenommen werden kann" (Bülow-Schramm / Braun 2013: 3). Insofern wird die AHELO-Initiative von deutschsprachigen Forschern zwar begrüßt, gleichzeitig aber auch auf eine Vielzahl an Problemen verwiesen, die es noch zu lösen gilt (Braun et al. 2013).

Die zweite große Forschungsinitiative, auf nationaler Ebene, ist das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor", in dessen Rahmen sich seit 2011 über 20 Forschungsprojekte der Entwicklung von Kompetenzmodellen und -messinstrumenten in fachspezifischen, aber auch fächerübergreifenden Domänen widmen. Einen Überblick geben Zlatkin-Troitschanskaja et al. (2012).[7] Eine ausführliche Diskussion der Ergebnisse dieser Forschungsprojekte kann aus Platzgründen hier nicht erfolgen und wäre auch verfrüht, da die Projekte noch nicht abgeschlossen sind und nur teilweise Publikationen vorliegen. Beispiele sind etwa ein Projekt zu „Kompetenzmodellierung und Kompetenzerfassung mathematischer Kompetenz als Teilkompetenz in den Ingenieurwissenschaften", eines zu „Kompetenzen Studierender im Umgang mit wissenschaftlicher Originalliteratur" oder eines zu „Modellierung und Messung wirtschaftswissenschaftlicher Fachkompetenz bei Studierenden bzw. Hochschulabsolventen". Im Bereich der Sozialwissenschaften widmet sich ein Projekt der „Modellierung und Messung wissenschaftlicher Kompetenz in sozialwissenschaftlichen Fächern", darunter Psychologie, Politikwissenschaft und Soziologie, wobei auch hier noch keine Ergebnisse publiziert sind.[8]

Für einige Fächer wurden aber bereits Kompetenzmodelle und -messinstrumente vorgestellt und hinsichtlich ihrer Validität diskutiert, was die grundsätzliche Gangbarkeit der Forschungsvorhaben demonstriert. Zum Beispiel präsentieren Kleickmann et al. (2013) Skalen zur Messung des bildungswissenschaftlichen Wissens von angehenden Lehrern und Riese / Reinhold

6 Ein weiteres Problem bei der übermäßigen Durchführung von Modellgeltungstests ist die „α-Überhöhung" (Kubinger / Draxler 2007) – bei dem üblichen Signifikanzniveau von 5% ist durchschnittlich einer von 20 Tests fälschlicherweise signifikant.

7 Siehe auch www.kompetenzen-im-hochschulsektor.de.

8 Die im vorliegenden Artikel präsentierte Pilotstudie steht in keiner Beziehung zu diesem Projekt.

(2012) zum Professionswissen (fachphysikalisches Wissen, physikdidaktisches Wissen, allgemeines pädagogisches Wissen) von angehenden Physiklehrern. Förster / Zlatkin-Troitschanskaia (2010) und Förster et al. (2012) stellen ein Instrument zur Messung wirtschaftswissenschaftlicher Fachkompetenzen vor. Obwohl also mittlerweile die zitierten und auch noch andere Modelle und Messinstrumente für eine Kompetenzdiagnostik im Hochschulbereich vorliegen, muss betont werden, dass sich diese nach wie vor in der Erprobungsphase befinden und in der Regel nur explorativ für inhaltliche bzw. evaluatorische Analysen eingesetzt wurden. Etwas weiter fortgeschritten sind diesbezüglich die Studien TEDS-M und TEDS-LT („Teacher Education and Development Study"„Mathematics" bzw. „Learning to Teach"). Im Rahmen von TEDS-M wurde in 16 Ländern unter mehr als 24.000 Lehramtsstudierenden deren mathematisches, mathematikdidaktisches und allgemeinpädagogisches Professionswissen erhoben; umfangreiche Ergebnisdarstellungen finden sich in Tatto et al. (2012) und Blömeke et al. (2010). TEDS-LT, die in Deutschland unter ca. 1.800 Studierenden durchgeführt wurde, fokussiert die fachlichen, fachdidaktischen und pädagogischen Kompetenzen von angehenden Deutsch-, Mathematik- und Englischlehrern (Blömeke et al. 2011; Blömeke et al. 2013).

Alles in allem verweisen die Forschungsaktivitäten einerseits auf die *Relevanz* einer Kompetenzdiagnostik im tertiären Bildungssektor und andererseits auf die generelle *Machbarkeit* entsprechender Vorhaben. Inwieweit sich die entwickelten Modelle und Skalen dann für die Hochschulevaluation oder inhaltliche Analysen zu Fragen des Bildungserwerbs und zu Bildungsrenditen eignen, muss sich erst noch zeigen – wobei dies letztlich eine empirische Frage ist und die TEDS-Studien in eine ermutigende Richtung weisen.[9]

3. Ein Pilotprojekt zur Messung soziologischer Kompetenzen

Die durchgeführte Pilotstudie hat sich explorativ der Frage nach der Definition und Messung soziologischer Kompetenzen gewidmet. Hierzu wurde versucht, ein Modell soziologischer Kompetenzen zu formulieren und dieses einer empirischen Messung zugänglich zu machen. Die folgenden Abschnitte erläutern die diesbezüglich getroffenen Entscheidungen.

3.1 Soziologische Kompetenzen im Modell

Konzentriert man sich zunächst auf fachliche und methodische soziologische Kompetenzen („Sachkompetenzen"), gilt es v.a. festzulegen, welche inhaltlichen soziologischen Dimensionen relevant sind. Zur Klärung dieser Frage sind grundsätzlich mehrere Herangehensweisen möglich, etwa die Analyse von Curricula und Modulhandbüchern soziologischer Studiengänge, die Sichtung von Lehrbüchern oder Veranstaltungsplänen, das Führen von Experteninterviews mit anerkannten Fachvertretern, oder ganz einfach Festlegungen, Empfehlungen oder Beschlüsse, die von den Forschern oder fachlichen Gremien, Verbänden (wie z.B. der DGS) oder Beiräten getroffen werden (können). Klar ist hierbei, dass eine solche Festlegung immer eine normative Entscheidung ist, die mehr oder weniger gut begründet und mehr oder weniger konsensuell, aber nicht „richtig" oder „falsch" sein kann. Im Fall der Soziologie wird diese generell schon anspruchsvolle Aufgabe noch dadurch erschwert, dass es sich um eine methodisch und vom Wissenschaftsverständnis her paradigmatisch heterogene Disziplin handelt, deren inhaltliche Problemstellungen bzw. Lehr- und Forschungsgegenstände (Stichwort „spezielle Soziologien") zudem sehr weit gespannt sind, nur wenig aufeinander aufbauen und oft auch völlig unabhängig voneinander betrieben werden („gering strukturierte Domäne"). Dazu

9 Der Vollständigkeit halber sei erwähnt, dass wir auf die vom Magazin DER SPIEGEL durchgeführte und plakativ als „Studentenpisa" etikettierte Erhebung nicht eingehen, da es sich in der durchgeführten Erhebung nicht um eine Kompetenzmessung, sondern um einen einfachen Wissenstest zur Allgemeinbildung handelt (vgl. für Näheres Trepte / Verbeet 2010).

gesellt sich noch der Umstand, dass die wechselseitige Akzeptanz der verschiedenen methodischen und theoretischen Strömungen als eher gering einzustufen und von gegenseitiger Ablehnung gekennzeichnet ist. Kurz gesagt, die Voraussetzungen, einen Konsens zu finden hinsichtlich der Inhalte, die in eine Definition soziologischer Kompetenzen eingehen sollen, dürften nicht die besten sein.

Wir sehen in dieser Ausgangskonstellation zwei Wege, die man einschlagen könnte. Eine erste Möglichkeit wäre, nach einem konsensfähigen kleinsten gemeinsamen Nenner an „soziologischen Basiskompetenzen“ zu suchen, die unabhängig von der methodischen und/oder theoretischen Ausrichtung zur Kompetenz von Soziologinnen und Soziologen zählen und eine Art Basisstandard identifizieren, der sich auch in der grundständigen Lehre in den meisten deutschen Soziologiestudiengängen wiederfinden dürfte. Dazu könnte beispielsweise zählen, schon einmal von Max Weber gehört zu haben oder anhand einer Situationsdarstellung einen Rollenkonflikt erkennen und benennen zu können. Damit ginge natürlich einher, dass z.B. auf die Integration vieler spezieller Soziologien oder selten benutzter Methoden oder Theorien in das Kompetenzmodell verzichtet werden müsste. Der zweite Weg wäre, von vornherein auf das Ziel eines integrativen Kompetenzmodells für „die Soziologie“ zu verzichten und stattdessen separate Kompetenzmodelle für enger abgegrenzte inhaltliche und methodische Domänen zu formulieren. Dies könnte so weit getrieben werden, dass man nur Kompetenzmodelle für einzelne Lehrveranstaltungen akzeptiert. Natürlich würde eine solche Vorgehensweise zu Lasten der Vergleichbarkeit der Kompetenzmaße und deren Einsetzbarkeit in hochschulvergleichenden Evaluationen gehen.

In der Pilotstudie wurde der erste Weg gewählt und die Absicht verfolgt, einen Pool an soziologischen Basiskompetenzen abzugrenzen, die in den meisten deutschsprachigen Soziologiestudiengängen vermittelt werden dürften. Für die Modellierung der Sachkompetenzen (also fachliche und methodische Kompetenzen) wurden fünf inhaltliche Dimensionen unterschieden: soziologische Begriffe, Theorien und Klassiker (bzw. Ideengeschichte), Sozialstrukturanalyse, quantitative Methoden, qualitative Methoden, sowie Infrastrukturkompetenz. Letztere bezeichnet methodische und praktische Kenntnisse und Fähigkeiten, die sich beispielsweise auf die Kenntnis von soziologischen Fachzeitschriften, Datenbanken, Geldgebern für Drittmitteleinwerbungen, wichtige Institutionen wie die DGS usw. beziehen. Die Wahl dieser Dimensionen stützte sich auf die Konsultation von Modulhandbüchern soziologischer Studiengänge, einführende Lehrbücher sowie die Empfehlungen der DGS zur Gestaltung soziologischer BA- und MA-Studiengänge von 2005 (DGS 2005).[10] Sicherlich müsste diese Festlegung in einem künftigen größeren Forschungsprojekt systematischer und begründeter erfolgen. Für jede dieser Subdimensionen wurden sodann konkrete Items entwickelt, die kondensierten kognitiven Prozessen entsprechen, und zwar kennen / verstehen, anwenden / interpretieren, beurteilen / auswählen / gestalten, gemäß der überarbeiteten Version der Bloom-Taxonomie.[11]

Eine wichtige Frage bei der Itementwicklung ist die „Bewertung“ der Testitems, also die jeweilige Entscheidung, welche Antwort auf eine Frage als „richtig“ gewertet wird. Während Items zu den unteren Bloom-Stufen – etwa einfache Wissensfragen mit geschlossenem Antwortformat – noch einfach zu entwerfen und hinsichtlich ihrer Bewertung unstrittig sind, ist

10 Für den BA empfiehlt das DGS-Papier eine Aufteilung von 25% der Arbeitsstunden auf „Soziologisches Denken / Theorien“, 20% „Methoden / Lehrforschung“, 10% Sozialstruktur, 30% „spezielle Soziologien und Vertiefungsgebiet“ und 15% BA-Arbeit. In der Pilotstudie wurde im Sinne der Fokussierung auf studiengangs-/hochschulübergreifende Basiskompetenzen auf die Integration von speziellen Soziologien zu Gunsten einer stärkeren Gewichtung der Sozialstruktur verzichtet.

11 Die Testitems wurden im Rahmen des Lehrforschungsprojekts von den teilnehmenden Studierenden und den beiden Autoren entwickelt.

es oft schwer, bei Evaluations- oder Anwendungsfragen – und generell bei offenen Fragen – zu entscheiden, welche Antwort als richtig zu werten ist und anhand welcher Kriterien dies festgelegt werden soll.

Insgesamt ist die Formulierung eines Kompetenzstrukturmodells und die Entwicklung von guten Testitems sicherlich die schwierigste Aufgabe, wenn in Zukunft eine standardisierte Kompetenzmessung in der Soziologie stattfinden soll. In künftigen Forschungsprojekten wäre vor allem in diesem Bereich noch intensive Grundlagenarbeit zu leisten. Auf die Ausgestaltung der Testitems in der Pilotstudie gehen wir im nun folgenden Abschnitt ein.

3.2 Fragebogen- und Erhebungsdesign

Die Befragung wurde als schriftliche Befragung konzipiert; der Fragebogen sollte von Soziologiestudierenden in Lehrveranstaltungen in einem Zeitrahmen von 45 Minuten bewältigt werden. Nachdem in einem Pretest (N = 37) die durchschnittlich benötigte Zeit für ein Testitem abgeschätzt wurde, wurde deren Zahl im Fragebogen auf 35 festgelegt. Zusätzlich wurde ein Testheftdesign gewählt, indem die Items in acht Blöcke à fünf Fragen aufgeteilt wurden und die Auswahl und Reihenfolge der Blöcke in sechs verschiedenen Fragebogenversionen bzw. Testheften variiert wurde. Dadurch konnten zum einen 40 und nicht nur 35 Items getestet werden, zum anderen können mögliche Halo-Effekte durch die Abfolge der Fragen im Fragebogen kontrolliert werden. Da im Vorfeld der Erhebung unklar war, wie viele Studierende den Fragebogen ausfüllen würden und um diesbezüglich möglichen Fallzahlproblemen begegnen zu können, waren 25 Testitems (fünf Blöcke für die fünf Dimensionen Begriffe / Theorien / Klassiker, Sozialstruktur, quantitative Methoden, qualitative Methoden, Infrastruktur) in jeder Fragebogenversion enthalten und wurden von jedem Studierenden beantwortet. Von den 15 weiteren Items wurden pro Testheft jeweils nur zehn Items (zwei Blöcke) bearbeitet.

Die 40 Testitems lassen sich nach einer Typologie von Jonkisz et al. (2012: 39) in 16 Fragen mit freiem Antwortformat und 24 Fragen mit gebundenem Antwortformat unterscheiden. Unter den Aufgaben mit freiem Antwortformat waren neun Items Ergänzungsaufgaben und sieben Items „Kurzaufsatzaufgaben“. Von den Items mit gebundenem Antwortformat waren sechs Zuordnungs- oder Umordnungsaufgaben, 16 Single-Choice-Aufgaben und zwei Multiple-Choice-Aufgaben. Abbildung 3 zeigt drei Beispielfragen aus dem Fragebogen. Die erste Frage soll die Dimension „Begriffe / Theorien / Klassiker“ messen und ist ein Beispiel für eine „Kurzaufsatzaufgabe“, die dem kognitiven Prozess „Gestalten“ entspricht. Die zweite abgebildete Frage entstammt der Dimension „qualitative Methoden“, ist eine Single Choice-Aufgabe und lässt sich auf der adaptierten Bloom-Taxonomie unter „Erinnern / Verstehen“ einordnen. Das letzte Fragebeispiel ist eine Ergänzungsfrage aus dem Bereich „quantitative Methoden“ und lässt sich mit den Bloom’schen Begriffen als Fähigkeit, „prozedurales Wissen anzuwenden“ beschreiben.

Abbildung 3: Beispielitems

FG1 **Überlegen Sie sich ein Beispiel für einen „Inter-Rollenkonflikt“, in dem ein Professor vorkommt und notieren Sie es unten.**

Antwort: _____

_____

_____

_____

FD5 **Welche soziologische Forschungsmethode beschreibt der folgende Satz (Geertz 1985: 38)?**

„Wir reden mit dem Bauern auf dem Reisfeld oder mit der Frau auf dem Markt, weitgehend ohne strukturierten Fragenkatalog und nach einer Methode, bei der eins zum anderen und alles zu allem führt; wir tun dies in der Sprache der Einheimischen, über eine längere Zeitspanne hinweg, und beobachten dabei fortwährend aus nächster Nähe ihr Verhalten.“

☐ 1 Objektive Hermeneutik
☐ 2 Inhaltsanalyse
☐ 3 Ethnographie
☐ 4 Triangulation

FC4 **Berechnen Sie Modus, Median und arithmetisches Mittel für diese Gruppe von Studenten:**

Student:	**1**	**2**	**3**	**4**	**5**
Alter:	**20**	**20**	**25**	**30**	**35**

a) Modus: _____

b) Median: _____

c) Arithmetisches Mittel: _____

Anhand der drei Fragebeispiele lassen sich auch die oben angesprochenen Probleme nachvollziehen: Die erste Frage beinhaltet zumindest tendenziell ein Bewertungsproblem. Die teils ausführlichen Antworten der Studierenden müssen hinsichtlich ihrer „Richtigkeit“ beurteilt werden; hierzu ist es nötig, ex ante Bewertungskriterien zu formulieren, die dann von den Vercodern reliabel angewendet werden müssen. In der Pilotstudie wurde dies im Seminarverbund umgesetzt, wobei die Problematik der Intercoder-Reliabilität dadurch entschärft wurde, dass die Items im Seminarverbund „live und vor Ort“ kodiert wurden, so dass eine große Transparenz hergestellt und bei strittigen Fällen gemeinsam entschieden werden konnte. Das zweite Beispiel der Single-Choice-Frage beinhaltet das Problem einer Ratewahrscheinlichkeit. Auch bei rein zufälligem Ankreuzen sind bei hinreichend großer Fallzahl 25% der Antworten korrekt. Derartigen Problemen kann im Rahmen der Itemanalyse durch die Modellierung von Rateparametern begegnet werden, deren Schätzung allerdings höhere Fallzahlen als die hier realisierten erfordert; eine unkompliziertere Variante, die auch in der hier präsentierten Studie eingesetzt wurde, ist, die Testpersonen deutlich darauf hinzuweisen, dass Fragen, deren Antwort man nicht kennt, leer gelassen werden sollen. Das dritte Fragebeispiel schließlich zeigt die Problematik auf, wie mit Items umgegangen werden soll, die mehrere Unteraufgaben beinhalten. Angebracht wären für solche Items mehrstufige IRT-Modelle wie das Partial Credit Modell (Masters 1982) oder das Graded Response Modell (Samejima 1969), für deren Schät-

zung allerdings höhere Fallzahlen als die hier realisierten nötig wären.[12] In der vorliegenden Studie wurden mehrstufige Items binär kodiert und, da sich die Testitems insgesamt als etwas zu schwer herausstellten, eine „milde" Bewertung gewählt und Mehr-Item-Aufgaben in der Regel als richtig gewertet, wenn mindestens die Hälfte der Unteraufgaben richtig gelöst wurden.

Neben den Testitems zur Messung soziologischer Kompetenz enthielt der Fragebogen weitere Module. In einem ersten Modul wurden Fragen zum Studium gestellt (z.B. Studiengang, Fachsemester, Klausur- und Abiturnoten). Außerdem war ein Modul mit Fragen zur Selbstwirksamkeitserwartung (Jerusalem / Schwarzer 2012) als Operationalisierung für das Selbstkompetenz-Konstrukt von Roth (1971) Gegenstand des ersten Teils des Fragebogens. Daraufhin folgten die jeweils 35 Testitems zur Messung soziologischer Kompetenz. Der letzte Teil des Fragebogens umfasste ein Modul mit Fragen zum Mainzer Institut für Soziologie und zur Studiumsmotivation, und eines zur Soziodemographie. Zwei methodische Fragen bildeten den Abschluss; zum einen wurde eine subjektive Einschätzung abgefragt, wie viel Mühe der oder die Studierende sich beim Ausfüllen der Testitems gegeben hat, zum anderen wurde ein personenspezifischer Code aus mehreren privaten Angaben (erste Ziffer des Geburtstags, erster Buchstabe des Vornamens der Mutter usw.) erfragt, um Doppelbefragungen identifizieren zu können.

3.3 Ablauf der Erhebung, Datenbasis und Variablen

Die Erhebung fand in den ersten Semesterwochen des Wintersemesters 2012/2013 in Lehrveranstaltungen des Mainzer Instituts für Soziologie statt. Die Befragung wurde durch das Leitungsgremium des Instituts offiziell unterstützt, was sich positiv auf die Kooperation der Lehrenden und Studierenden auswirkte. In der jeweiligen Lehrveranstaltung stellten im Projekt mitarbeitende Studierende das Ziel der Erhebung kurz vor, verwiesen auf Freiwilligkeit und Anonymität und betonten, dass Fragen unausgefüllt bleiben sollten, falls der oder die Studierende die Antwort nicht wusste. Ebenso wurde darauf verwiesen, dass insbesondere Studierende in den unteren Semestern viele Fragen nicht würden beantworten können, dass dies nicht zu vermeiden sei und man sich dadurch nicht entmutigen lassen sollte.[13] Sodann verteilten die studentischen Mitarbeiter die Fragebögen, wobei die sechs verschiedenen Fragebogenversionen randomisiert ausgeteilt wurden.

Aufgrund der schwankenden und insbesondere in Vorlesungen nicht genau bestimmbaren Teilnehmerzahl (gerade zu Beginn des Semesters) können keine zahlenmäßigen Angaben zur Grundgesamtheit und zum Ausfall (Unit-Nonresponse) gemacht werden. Angesichts des primär methodisch-explorativen Erkenntnisinteresses der Studie sind die geringeren Anforderungen, die an die Stichprobe gestellt wurden, unseres Erachtens aber vertretbar (siehe auch unten). Die Erhebung fand in 15 Lehrveranstaltungen aller Semester statt, die Zahl der ausgefüllten Fragebögen pro Kurs schwankte zwischen 5 und 138 Bögen. Insgesamt wurden 540 Fragebögen ausgefüllt. Von diesen wurden anhand des fünfstelligen persönlichen Codes am Ende des Fragebogens sowie der Variablen Geschlecht und Geburtsjahr 59 Dubletten identifiziert, die für die hier berichteten Analysen aus dem Datensatz entfernt wurden (beibehalten wurde immer der zuerst erhobene Fall).

12 Eine weitere Möglichkeit bestünde darin, jede Teilaufgabe als eigenständiges Item zu behandeln und in die statistische Analyse eingehen zu lassen. Dies kann jedoch dem Ziel der lokalen Unabhängigkeit der Testitems widersprechen.

13 Eine Lösung zur Vermeidung dieses Problems wäre ein adaptiver Test, bei dem im Testverlauf computergestützt die Schwierigkeit der Aufgaben an das Kompetenzniveau des Probanden angepasst wird. Hierfür wären kalibrierte Items nötig, deren Schwierigkeit bereits bekannt sein müsste.

Das Fazit zur erhebungstechnischen und praktischen Durchführbarkeit von vergleichbaren Erhebungen fällt positiv aus: Sowohl Lehrende, als auch Studierende kooperierten besser als gedacht, letztere füllten den Fragebogen größtenteils gewissenhaft aus. Unit-Nonresponse trat nur sporadisch auf und konnte recht einfach dadurch, dass die Befragung zu Beginn der jeweiligen Lehrveranstaltung und nicht am Ende durchgeführt wurde, vermieden werden.

4. Ergebnisse: Ein Indikator zur Messung soziologischer Kompetenzen

4.1 Ergebnisse der Skalierung

Für die Skalierung wurden zunächst alle 40 Testitems bewertet und binär kodiert, so dass jede Aufgabe entweder als richtig, falsch bzw. nicht gelöst, oder fehlend (Items der Substichprobe) gilt. Sodann wurde die Skalierung mit den IRT-Routinen der R-Pakete ltm (Rizopoulos 2012) und eRm (Mair et al. 2013) vorgenommen. Die Ergebnisse sind in Tabelle 1 dargestellt.

Tabelle 1: Ergebnisse zur Skalierung der Einzeldimensionen

	Beg./Th./Kl.	**Sozialstruktur**	**Quanti-Methoden**	**Quali-Methoden**	**Infrastruktur**
Erhobene Items	12	10	6	5	7
Ausgewählte Items	10	8	6	4	5
Gewähltes Modell	Rasch	Rasch	Rasch	Rasch	2PL
DIF θ-Split: z (df)	18,11 (8)	7,86 (7)	8,80 (4)	0,45 (3)	n. k.
DIF Geschlecht: z	9,16 (9)	8,75 (7)	1,99 (5)	7,22 (3)	17,68 **
Bootstrap-GoF: T	739,13	366,47	315,87	5,78	n. k.
Eindimensionalitätstest: Eigenwertdifferenz	0,11	0,01	n. k.	0,06	0,03
LR-Test 2PL: χ^2 (df)	34,23 (9) ***	8,25 (7)	n. k.	2,49 (3)	449,60 (4) ***

Erläuterungen: Der Bootstrap-Goodness of Fit-Test wurde mit 501 Datensätzen simuliert, der Eindimensionalitätstest mit 100 Monte-Carlo-Stichproben. „n. k.“ steht für nicht konvergierte Maximum-Likelihood-Schätzungen.

In der Tabelle sind mehrere Modelltests, die für jede Subskala durchgeführt wurden, zusammengefasst. Hierbei sind signifikante Testwerte grundsätzlich als unerwünscht zu interpretieren. Das für jede Subdimension geschätzte Modell wurde erstens auf unterschiedliche Schätzungen der Aufgabenparameter in verschiedenen Subgruppen des Datensatzes getestet (DIF). Hierdurch wird sichergestellt, dass die Items in diesen Subgruppen konsistent funktionieren und nicht für bestimmte Studierende leichter oder schwerer zu lösen sind. Der Test „θ-Split“ teilt die Stichprobe am Mittelwert der Skalenwerte θ und schätzt die Aufgabenparameter in beiden Gruppen. Die zwei Schätzungen werden sodann mit einem Likelihood-Ratio-Test gegeneinander auf signifikante Abweichungen getestet. Der DIF-Test nach Geschlecht funktioniert analog. Zweitens wurde ein Bootstrap-Goodness of Fit-Test durchgeführt, der auf Abweichungen der laut Rasch-Modell erwarteten Verteilung der Antwortmuster mit den beobachteten vergleicht (Strobl 2012: 47). Drittens wurde ein Eindimensionalitätstest durchgeführt, der den zweiten Eigenwert der tetrachorischen Korrelationsmatrix der Items mit aus der Modellschätzung simulierten Korrelationen vergleicht und die Abweichung auf Signifikanz testet. Zuletzt wurde viertens jeweils ein 2PL-Modell berechnet und dessen Anpassung an die Daten mit jener des betreffenden Rasch-Modells verglichen. Die Testprozeduren wurden wiederholt und sukzessive Items ausgeschlossen, bis brauchbare Ergebnisse erreicht wurden.

Die Ergebnisse zeigen insgesamt zufriedenstellende Resultate. Allerdings traten an einigen Stellen Konvergenzprobleme der Maximum-Likelihood-Algorithmen auf; beispielsweise konnte für die Skala „Quantitative Methoden“ kein 2PL-Modell berechnet werden, für die Subskala „Infrastruktur“ konvergierte das Rasch-Modell nicht. Für die Subskala „Begriffe / Theorien / Klassiker“ wurde trotz eines signifikant (aber nicht erheblich) besseren 2PL-Mo-

dells nach einer Inspektion der Items und aufgrund der sämtlich nicht signifikanten weiteren Modelltests für das Rasch-Modell dennoch ein solches für die Skalierung herangezogen. Einzig für die Subskala „Infrastruktur" musste ein 2PL-Modell zugrunde gelegt werden.

Für jede Subskala wurden abschließend die Personenparameter, also der jeweilige Wert der latenten Variable für jeden Fall, abgespeichert. Eine Bayes-Routine des Programms ltm für R schätzt aufgrund der verfügbaren Items (Testheftdesign) hierbei für sämtliche Fälle im Datensatz den Wert der latenten Variable und deren Standardfehler – bei Personen, die einen Teil der Items nicht beantwortet haben, gehen entsprechend nur weniger Items in die Schätzung ein.

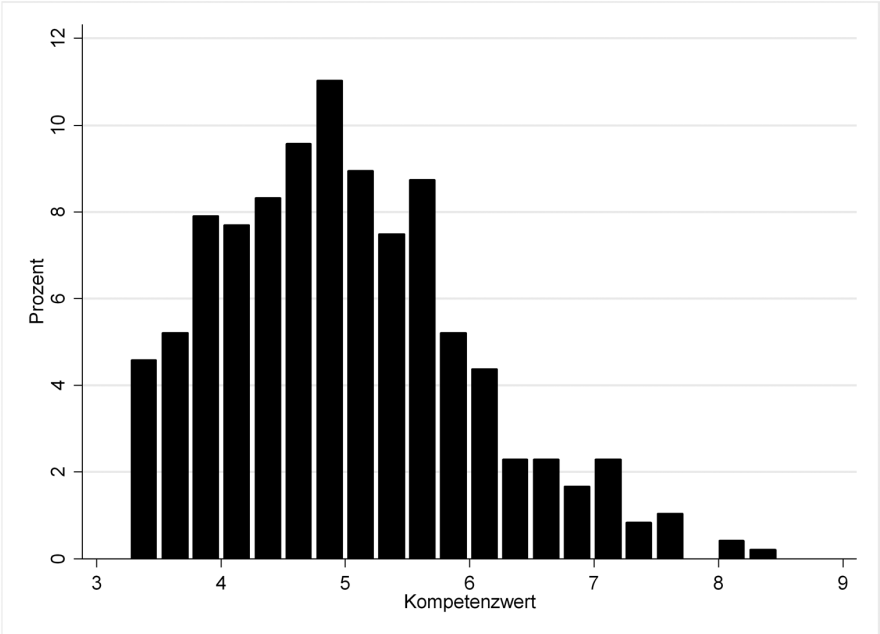
Um die fünf Subskalen für inhaltliche Analysen im nächsten Abschnitt zu einem „Globalindikator"[14] zusammenzufassen, wurde eine Hauptkomponentenanalyse durchgeführt, deren Ergebnisse auf eine eindimensionale Lösung hinweisen (der Eigenwert der zweiten Hauptkomponente beträgt nur noch 0,65). Die extrahierte Hauptkomponente erklärt gut 57% der Varianz, alle Items laden deutlich und hoch auf der Komponente. Der Reliabilitätskoeffizient (Cronbachs α) der fünf Skalenitems beträgt 0,81. Beide Ergebnisse rechtfertigen die Zusammenfassung der fünf Einzelindikatoren zu einer Variablen. Hierzu wurden die Faktorwerte nach der Regressionsmethode gespeichert und die resultierende Variable standardisiert (mit Mittelwert = 5 und Standardabweichung = 1).[15]

Abbildung 4 zeigt die Verteilung des Kompetenzindikators. Die Verteilung ist leicht rechtsschief, was als Hinweis dafür gewertet werden kann, dass der Test insgesamt für die hier untersuchte Stichprobe zu schwer war. Dieser Eindruck bestätigt sich auch, wenn man die Lösungshäufigkeiten der einzelnen Items betrachtet (hier nicht dokumentiert).

14 Dies ist nur eine mögliche Vorgehensweise. Alternativ könnten die fünf Subskalen auch einzeln in empirische Analysen eingehen.

15 Allerdings zeigt der Vergleich eines eindimensionalen Rasch- bzw. Strukturgleichungsmodells mit einem mehrdimensionalen eine signifikant bessere Anpassung des mehrdimensionalen Modells. In letzterem korrelieren die fünf Einzeldimensionen allerdings hoch mit Korrelationen zwischen 0,7 und 0,9, was u.E. die eindimensionale Behandlung der Skala rechtfertigt.

Abbildung 4: Verteilung des globalen Kompetenzindikators



Erläuterung: N = 481. Die Variable wurde auf einen Mittelwert von 5 und eine Standardabweichung von 1 standardisiert.

Insgesamt sind die vorgestellten Skalen zur Messung soziologischer Kompetenzen selbstverständlich bei weitem nicht ausgereift und, was die Itemanzahl angeht, stark ausbaufähig. Dennoch sehen wir die Ergebnisse dieser ersten sondierenden Annäherung an eine Kompetenzmessung in der Soziologie als ermutigend, anschluss- und ausbaufähig an. Die Skalierung der Items mit IRT-Verfahren funktionierte besser als erwartet.

4.2 Ausgewählte Analysen mit dem Kompetenzindikator

Im Folgenden werden Regressionsanalysen des Kompetenzindikators auf mehrere Einflussfaktoren präsentiert. Dies dient zum einen einer Konstruktvalidierung des Indikators: Wenn dieser tatsächlich „soziologische Kompetenzen“, die im Rahmen eines soziologischen Hochschulstudiums vermittelt werden, messen soll, so ist nachzuweisen, dass die Variable mit relevanten und getrennt erhobenen Außenkriterien korreliert. Zum anderen dienen die Analysen dazu, einige interessante inhaltliche Befunde aufzuzeigen, welche die Motivation und Relevanz einer Kompetenzmessung im Hochschulbereich nochmals unterstreichen sollen.

Für die Analysen wurden zunächst einige Hintergrundmerkmale aufbereitet. Tabelle 2 gibt eine Übersicht. Die Soziologiestudierenden in der Stichprobe haben im Durchschnitt gut drei Fachsemester Soziologie studiert, befinden sich also im Mittel im vierten Fachsemester; 37% der Befragten sind Nebenfachstudierende. Die Indikatoren „Selbstwirksamkeit“ und „intrinsische fachliche Motivation“ wurden als Operationalisierung der „Selbstkompetenz“ im Sinne Roths (1971, vgl. Abschnitt 2.1) erhoben. Selbstwirksamkeit soll „Überzeugungen subjektiver Kontrollierbarkeit bzw. Kompetenzerwartungen“ (Jerusalem / Schwarzer 2012) in studiums-

bezogenen Situationen erfassen. Konkret geht es um die subjektive Einschätzung, für wie kompetent und leistungsfähig sich Studierende in Prüfungssituationen halten. Die Variable ist ein Summenindex aus sieben (eindimensional ladenden) Einzelitems. Der Indikator „intrinsische fachliche Motivation" wurde von den Autoren entwickelt und bezieht sich auf intrinsisches Interesse am Fach Soziologie (z.B. zusätzlich zum Studium soziologische Texte lesen). Drei eindimensional ladende Items wurden zu einem Summenindex zusammengefasst. Die mittlere Abiturnote der Soziologiestudierenden liegt bei 2,5 mit einer Streuung von 0,5.[16] Gut die Hälfte der Befragten gibt an, neben dem Studium häufig oder laufend einer Erwerbstätigkeit nachzugehen. Als Indikator für die soziale Herkunft wurden die Bildung des Elternhauses und der Migrationshintergrund erhoben. Die Bildung der Eltern wurde aus Angaben zum allgemeinen und beruflichen Abschluss von Vater und Mutter kodiert, in Bildungsjahre approximiert und, falls Angaben zu beiden Eltern vorlagen, gemittelt. Als Durchschnitt in der Stichprobe ergibt sich ein Wert von 14,2 Jahren bei einer Standardabweichung von 2,6 Jahren. Der Indikator für Migrationshintergrund beruht auf Angaben zum Geburtsland der Studierenden und deren Eltern. Die Variable erhält den Wert eins, wenn entweder der Studierende selbst oder mindestens ein Elternteil nicht in Deutschland geboren ist. Nach dieser Berechnung haben 28% der Befragten einen Migrationshintergrund. Als weitere Kontrollvariable wurde die subjektive Einschätzung der Deutschkenntnisse abgefragt. Diese wird von den Befragten mit einem Mittelwert von 4,5 im Durchschnitt als gut bis sehr gut eingeschätzt. Die Auszählung der Geschlechtsvariable bestätigt mit 61% weiblichen Studierenden den überdurchschnittlichen Frauenanteil unter Studierenden in sozialwissenschaftlichen Studiengängen. Schließlich wird in die Regressionsanalysen eine Kontrollvariable „Mühe beim Ausfüllen der Testitems" eingehen, deren Mittelwert auf einer Skala von eins bis sieben bei 4,4 liegt.

Tabelle 2: Übersicht der verwendeten Variablen

Variable	Ausprägungen/Bemerkungen	MW	SA	N
Kompetenzindikator	Siehe Abschnitt 4.1	5	1	481
Fachsemester	Absolvierte Fachsemester in Soziologie [0…18]	3,36	2,94	476
Studiengang	0 = Soziologie BA Kernfach/Master/Magister/Diplom 1 = Nebenfach, Sonstiges	0,37		478
Selbstwirksamkeit	1 = gering bis 7 = hoch (Erläuterung siehe Text)	4,07	0,81	452
Intrinsische fachliche Motivation	1 = gering bis 7 = hoch (Erläuterung siehe Text)	4,34	1,21	466
Abiturnote	Gesamt-Abiturnote [1…5]	2,50	0,52	460
Erwerbstätigkeit neben dem Studium	1 = laufend/häufig, 0 = nie/gelegentlich	0,56		470
Bildung der Eltern	In Jahren (Erläuterung siehe Text)	14,24	2,62	447
Migrationshintergrund	1 = ja, 0 = nein (Erläuterung siehe Text)	0,28		467
Deutschkenntnisse	Subjektive Einschätzung von 1 = mangelhaft bis 5 = sehr gut	4,47	0,79	475
Geschlecht	1 = weiblich, 0 = männlich	0,61		472
Mühe beim Ausfüllen	Subjektive Einschätzung der gegebenen Mühe beim Bearbeiten der Testitems von 1 = wenig Mühe bis 7 = sehr große Mühe	4,40	1,57	473

Erläuterung: MW Mittelwert, SA Standardabweichung, N Fallzahl.

Die Struktur der fehlenden Werte (keine Angabe) zeigt wenige Leerzellen, keine Variable weist ein Übermaß an Item-Nonresponse auf. Um zu vermeiden, dass sich die Fallzahl bei einem listenweisen Fallausschluss in den Regressionsanalysen zu sehr reduziert, wurden die fehlenden Werte durch multiple Imputation imputiert. Hierbei wurde nach den Empfehlungen in StataCorp (2011) vorgegangen und aufgrund der Missing-Struktur mit „chained equations"

16 Eine Person hat hier eine Note von 5 angegeben – da es sich um nur einen Fall handelt und der Wert möglicherweise doch plausibel sein kann (Hochschulzugangsberechtigung auf anderen Wegen erworben), wurde der Fall nicht ausgeschlossen.

unter Heranziehung von 14 zusätzlichen Imputationsvariablen 20 Imputationsdatensätze erzeugt. Diese wurden sodann mit OLS-Regressionen für imputierte Daten analysiert. Die Ergebnisse zweier Modellschätzungen sind in Tabelle 3 wiedergegeben.[17]

Tabelle 3: Determinanten soziologischer Kompetenz (OLS-Regressionen mit multipel imputierten Daten)

	Modell 1	**Modell 2**
Fachsemester	0,361 ***	0,385 ***
	(0,032)	(0,031)
Fachsemester quadriert	-0,016 ***	-0,018 ***
	(0,003)	(0,003)
Studiengang (1 = Nebenfach)		-0,183 *
		(0,079)
Selbstwirksamkeit		0,064
		(0,046)
Intrinsische fachliche Motivation		0,087 **
		(0,031)
Abiturnote		-0,146 *
		(0,072)
Erwerbstätigkeit neben dem Studium (1 = ja)		0,019
		(0,072)
Bildung der Eltern in Jahren		-0,025 +
		(0,014)
Migrationshintergrund (1 = ja)		-0,308 ***
		(0,081)
Deutschkenntnisse		0,173 ***
		(0,045)
Geschlecht (1 = weiblich)		-0,248 **
		(0,075)
Mühe beim Ausfüllen		0,102 ***
		(0,023)
Konstante	4,112 ***	3,213 ***
	(0,071)	(0,435)
Korrigiertes Pseudo-R^2	0,345	0,477

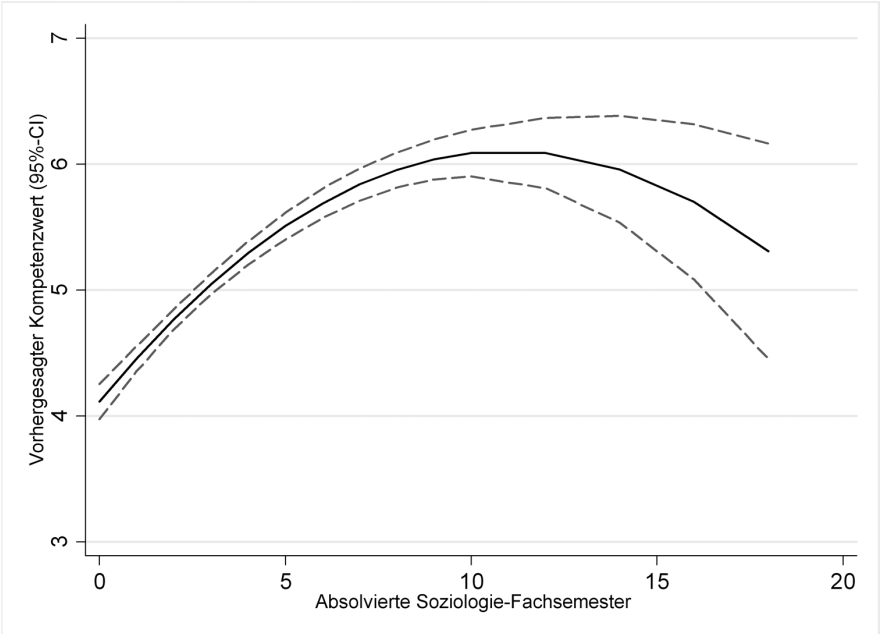
Erläuterung: Die abhängige Variable ist standardisiert mit einem Mittelwert von fünf und einer Standardabweichung von eins. Ausgewiesen sind unstandardisierte Regressionskoeffizienten und deren Standardfehler in Klammern. Fallzahl = 481. Signifikanzniveaus: +: $p < 0,1$; *: $p < 0,05$; **: $p < 0,01$; ***: $p < 0,001$.

Das erste Modell enthält ausschließlich das Fachsemester in Soziologie als Prädiktor – in Form eines linearen und quadratischen Terms. Der hochsignifikante Effekt erklärt bereits

17 Die Ergebnisse unterscheiden sich nicht wesentlich von (nicht dokumentierten) Regressionen, die mit listenweisem Fallausschluss durchgeführt wurden (N = 387).

rund 35% der Varianz des Kompetenzwertes, was für die Konstruktvalidität des Indikators spricht. In Abbildung 5 sind die aus der Modellschätzung vorhergesagten Kompetenzwerte gegen das Fachsemester abgetragen. Wie ersichtlich wird, zeigt sich bis zum etwa zehnten (absolvierten) Semester eine deutliche Kompetenzzunahme, die allerdings einen gewissen Sättigungseffekt oder abnehmenden Grenznutzen der Studiendauer im Studiumsverlauf aufweist. Etwa ab dem elften Fachsemester sinken die Werte wieder, wobei der deutliche Abfall der Kurve nicht überbewertet werden sollte, da die Fallzahlen hier gering sind und die Schätzung von Ausreißern beeinflusst sein kann. Inhaltlich gesehen ist aber eine abfallende Kurve in sehr hohen Semestern nicht unplausibel, da höchstwahrscheinlich ein Selektionseffekt in der Form vorliegt, dass „gute" Studenten nach zehn Semestern bereits ihren Abschluss gemacht haben oder zumindest nicht mehr in Lehrveranstaltungen, in denen die Erhebung durchgeführt wurde, präsent sind.

Abbildung 5: Vorhergesagte Kompetenzwerte nach Fachsemester



Erläuterung: Vorhergesagte Werte des Kompetenzindikators aufgrund von Modell 1, Tabelle 3.

Modell 2 in Tabelle 3 enthält ein ausführlicheres Regressionsmodell mit mehreren Prädiktoren, von denen ein Einfluss auf den Kompetenzindikator erwartet werden kann. Nebenfachstudierende schneiden etwas schlechter ab als Hauptfachstudierende. Unter den Indikatoren für „Selbstkompetenz“ hat nur die intrinsische Motivation für soziologische Themen einen (positiven) Einfluss auf den erreichten Kompetenzwert; der partielle Effekt der Selbstwirksamkeit ist nicht signifikant. Weiterhin zeigt sich – unter Kontrolle sämtlicher anderer Prädiktoren – ein signifikanter Einfluss der Abiturnote auf den Kompetenzerwerb: Je besser (also je niedriger) die Abiturnote, desto höher der erreichte Kompetenzwert. Eine Erwerbstätigkeit während des Studiums hat interessanterweise keinen Einfluss auf den Kompetenzerwerb. Sollte sich dieser Befund in weiteren Studien bestätigen, wäre dies ein wichtiges Argument für die Debatte, ob Jobben neben dem Studium den Studienerfolg – mutmaßlich aus Zeitgründen – beeinträchtigt.

Als Operationalisierung für die soziale Herkunft der Studierenden, und um mögliche primäre Herkunftseffekte (Boudon 1979) auf den Bildungserfolg – hier im Hochschulbereich – zu erfassen, wurde die Bildung der Eltern in das Modell integriert. Der Effekt ist nur auf dem 10%-Niveau signifikant und negativ, d.h. eine höhere Bildung des Elternhauses geht ceteris paribus tendenziell mit einem niedrigeren Kompetenzwert einher. In nicht dokumentierten weiteren Analysen hat sich allerdings gezeigt, dass der Effekt eher instabil ist und nicht überinterpretiert werden sollte. Deutlich und hochsignifikant sind hingegen die Effekte eines Migrationshintergrundes, der sich negativ auf den Kompetenzwert auswirkt, sowie, unabhängig davon, der Deutschkenntnisse, die sich positiv auswirken. Dieser (ungleichheits-)soziologisch relevante Befund wäre in künftigen Studien eingehender zu untersuchen, er demonstriert aber die Relevanz einer Kompetenzmessung im Hochschulbereich für inhaltliche Fragestellungen.

Des Weiteren zeigt sich ein hochsignifikanter Geschlechtseffekt in der Form, dass weibliche Studierende ceteris paribus etwa eine viertel Standardabweichung schlechter abschneiden als männliche Studierende. Dieser Befund ist insofern von Bedeutung, als er der Diagnose eines in jüngerer Zeit besseren Abschneidens von Frauen im deutschen Bildungssystem entgegensteht (Blossfeld et al. 2009; Hadjar 2011). Hingegen hat sich der Effekt auch in anderen Kompetenzerhebungen im Hochschulbereich immer wieder gezeigt (Förster et al. 2012; Walstad / Robson 1997; Zlatkin-Troitschanskaia et al. 2012; vgl. auch den Überblick bei Spiel et al. 2008: 68ff und die dort zitierte Literatur). Grundsätzlich steht jedoch zur Debatte, ob der Geschlechtseffekt tatsächlich unterschiedliche Kompetenzniveaus abbildet oder auf geschlechtsspezifisch variierendem Beantwortungsverhalten in Leistungstests beruht, also letztlich ein Artefakt darstellt. Demnach könnten laut Förster et al. (2012) und Spiel et al. (2008) Frauen eine geringere Rateprävalenz oder höhere Risikoaversion als Männer an den Tag legen, also bei Nichtwissen Aufgaben eher als Männer nicht beantworten. Walstad / Robson (1997) weisen für einen Test zur Messung ökonomischer Kompetenzen geschlechtsspezifisches DIF (differential item functioning) nach; Strobl / Kopf (2010) für einen Allgemeinbildungstest unter Studierenden. Beide Studien kommen jedoch zu dem Ergebnis, dass geschlechtsspezifische Unterschiede in den Testergebnissen nicht vollständig auf DIF zurückzuführen sind. Für die Interpretation als Artefakt spricht in unserer Studie die Tatsache, dass für 91 zu beantwortete Teilfragen von 40 Items der Mittelwert der Zahl nicht beantworteter Fragen für Frauen bei 43, für Männer bei 35 liegt ($T = 4{,}55$, $p < 0{,}001$). Dagegen sprechen die Ergebnisse zur Skalierung (Tabelle 1), nach denen sich für vier von fünf Subskalen kein DIF nach Geschlecht nachweisen ließ. Sicher ist auch, dass sich der bivariat noch ausgeprägtere Geschlechtseffekt (nicht dokumentiert) auch dadurch erklärt, dass weibliche Studierende eine signifikant niedrigere intrinsische Motivation und Selbstwirksamkeit als männliche Studierende aufweisen ($p < 0{,}001$ bzw. $p < 0{,}01$, nicht dokumentiert). Alles in allem sollte die Problematik in weiteren Analysen und Studien genauer untersucht werden[18] – die Befunde unterstreichen jedoch, dass es wichtig ist, in Analysen zum Bildungserwerb nicht nur konventionelle Maße wie Noten oder Zertifikate zu betrachten, sondern auch Kompetenzmaße heranzuziehen.[19] Der letzte Effekt des Regressionsmodells bezieht sich auf die methodische Kontrollvariable der Mühe beim Beantworten der Testitems, welche den erwarteten positiven Effekt zeigt.

Zusammengenommen erklären die unabhängigen Variablen knapp 50% der Variation des Kompetenzindikators, was angesichts der wenigen untersuchten Effekte als hoch betrachtet werden kann. Zum einen spricht dies für die Konstruktvalidität des Kompetenzindikators, zum anderen unterstreicht dies die Wichtigkeit und Relevanz von Kompetenzmessungen im Hochschulbereich.

5. Schlussfolgerungen und Ausblick

Der vorliegende Beitrag hat sich angesichts eines grundsätzlichen Bedarfs einer Kompetenzdiagnostik im Hochschulbereich und aktueller Forschungsanstrengungen in anderen Fächern der Hochschullandschaft der Frage nach einer Kompetenzmessung in der Soziologie gewidmet, also der Frage nach Natur und Inhalt sowie der Messung von „soziologischen Kompetenzen“. Zentral war hierbei die Präsentation von Erkenntnissen und Ergebnissen einer ersten empirischen Pilotstudie zum Thema.

18 Interessant wäre es z.B. zu untersuchen, ob der Geschlechtseffekt in verschiedenen Subdimensionen der Kompetenz variiert. Wegen der teils geringen Zahl der Items für die einzelnen Subdimensionen im vorliegenden Test haben wir darauf verzichtet, entsprechende Analysen durchzuführen.

19 Am Rande sei in diesem Zusammenhang bemerkt, dass weibliche Studierende in der Stichprobe eine um 0,19 Notenpunkte bessere Abiturnote haben als männliche Studierende ($p < 0{,}001$).

Als ermutigende Erkenntnis aus der Pilotstudie ist zunächst die praktische Durchführbarkeit zu nennen: Die Erhebung von Kompetenzitems im Rahmen soziologischer Lehrveranstaltungen gestaltete sich unproblematisch – die Studierenden bearbeiteten den Fragebogen in der Regel aufgeschlossen und konzentriert; Kritik, Zweifel oder andere Ängste traten nicht auf. Hierbei erwies sich auch die Testdauer – ca. 45 Minuten für den ganzen Fragebogen – als sinnvolle Orientierung für künftige Erhebungen zum Thema. Als weitere Aktiva der explorativen Studie sind die gute Skalierbarkeit der Kompetenzitems mit einfachen IRT-Modellen sowie der Gehalt des hier herangezogenen globalen Kompetenzindikators für inhaltliche empirische Analysen zu nennen. Allein der starke Zusammenhang mit den absolvierten Fachsemestern spricht für eine hohe Konstruktvalidität des Indikators.

Die wesentlichen Verkürzungen der hier präsentierten Studie sind nicht zuletzt den geringen Ressourcen einer ersten explorativen Studie geschuldet: Zum einen müsste mehr Energie in die Formulierung eines Kompetenzmodells investiert werden, welches im Idealfall auch noch in der soziologischen Community einen gewissen Grad an Konsens genießen sollte. Gleichzeitig wäre das Modell in eine größere Zahl an Items umzusetzen, als hier geleistet werden konnte. Dies erfordert, da die Zahl der Testaufgaben in einem Fragebogen beschränkt ist, auch eine deutlich höhere Fallzahl und eine Umsetzung in einem Testheftdesign.

Für die Zukunft bleibt also viel zu tun, wobei wir als zentrales Ergebnis der Pilotstudie die Erkenntnis sehen, dass die Probleme letztlich Ressourcenprobleme sind und – wie auch die Forschung zu Kompetenzdiagnostik in anderen Fächern zeigt – mit entsprechendem Aufwand lösbar sind. Sicherlich wäre das Vorhaben, eine Kompetenzmessung für die Soziologie zu entwickeln, ein Projekt für mehrere Jahre und eine größere Forschergruppe – hilfreich wäre dabei in jedem Fall auch institutionelle Unterstützung von soziologischen Verbänden, Gremien und Instituten. Gerade die Deutsche Gesellschaft für Soziologie hat mit ihrer Kritik am CHE-Ranking den Bedarf für validere, objektivere Messinstrumente im Bereich der Hochschulevaluation deutlich gemacht. Die Vision wäre in diesem Zusammenhang eine standardisierte, über Hochschulen vergleichbare und wiederholt durchführbare Erhebung soziologischer Kompetenzen im Stil der PISA-Studien. Dabei ist klar, dass das grundsätzliche Problem der geringen Strukturiertheit der Soziologie ein großes Problem darstellt. Eine weitere Möglichkeit ist, nach höher strukturierten Subdomänen im Bereich des Faches zu suchen und hierzu separate Kompetenzmodelle und -messungen durchzuführen. In einer gerade laufenden zweiten Pilotstudie eruieren wir dementsprechend Möglichkeiten einer Kompetenzmessung in den quantitativen Methoden der empirischen Sozialforschung.

Literatur

Anderson, Lorin W. / David R. Krathwohl (Hrsg.) (2001): A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom’s Taxonomy of Educational Objectives, New York / NY.

Baumert, Jürgen / Wilfried Bos / Rainer Watermann (1998): TIMSS/III. Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich. Zusammenfassung deskriptiver Ergebnisse, Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, Jürgen / Cordula Artelt / Eckhard Klieme / Johanna Neubrand / Manfred Prenzel / Ulrich Schiefele / Wolfgang Schneider / Klaus-Jürgen Tillmann / Manfred Weiß (Hrsg.) (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich, Opladen.

Baumert, Jürgen / Mareike Kunter (2006): Stichwort: Professionelle Kompetenz von Lehrkräften, in: Zeitschrift für Erziehungswissenschaft 9, S. 469-520.

Birnbaum, Allan (1968): Some Latent Trait Models, in: Frederic M. Lord / Melvin R. Novick (Hrsg.), Statistical Theories of Mental Test Scores, Reading / MA, S. 395-479.

Blömeke, Sigrid (2013): Ja, aber... – Lehren aus TEDS-M und KoKoHs für eine Teilnahme an AHELO, in: Edith Braun / André Donk / Margret Bülow-Schramm (Hrsg.), AHELO Goes Germany? Dokumentation des GfHf- & HIS-HF-Workshops, HIS: Forum Hochschule 2 / 2013, Hannover, S. 13-20.

Blömeke, Sigrid / Albert Bremerich-Vos / Helga Haudeck / Gabriele Kaiser / Günter Nold / Knut Schwippert / Heiner Willenberg (Hrsg.) (2011): Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen. Erste Ergebnisse aus TEDS-LT, Münster.

Blömeke, Sigrid / Albert Bremerich-Vos / Gabriele Kaiser / Günter Nold / Helga Haudeck / Jörg-U. Keßler / Knut Schwippert (Hrsg.) (2013): Professionelle Kompetenzen im Studienverlauf. Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrerausbildung aus TEDS-LT, Münster.

Blömeke, Sigrid / Gabriele Kaiser / Rainer Lehmann (Hrsg.) (2010): Professionelle Kompetenz an und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich, Münster.

Bloom, Benjamin S. (Hrsg.) (1956): Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain, New York / NY.

Blossfeld, Hans-Peter / Wolfgang Bos / Bettina Hannover / Dieter Lenzen / Detlef Müller-Böling / Manfred Prenzel / Ludger Wößmann (Hrsg.) (2009): Geschlechterdifferenzen im Bildungssystem. Jahresgutachten 2009, Wiesbaden.

Böhme, Katrin / Dirk Richter / Petra Stanat / Hans Anand Pant / Olaf Köller (2012): Die länderübergreifenden Bildungsstandards in Deutschland, in: Petra Stanat / Hans Anand Pant / Katrin Böhme / Dirk Richter (Hrsg.), Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011, Münster u.a., S. 11-18.

Bos, Wilfried Hornberg, Sabine / Karl-Heinz Arnold / Gabriele Faust / Lilian Fried / Eva-Maria Lankes / Knut Schwippert / Renate Valtin (Hrsg.) (2007): IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich, Münster.

Boudon, Raymond (1979): L'inégalité des chances, Paris.

Braun, Edith / André Donk / Margret Bülow-Schramm (Hrsg.) (2013): AHELO Goes Germany? Dokumentation des GfHf- & HIS-HF-Workshops, HIS: Forum Hochschule 2 / 2013, Hannover.

Braun, Edith / Burkhard Gusy / Bernhard Leidner / Bettina Hannover (2008): Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp), in: Diagnostica 54, S. 30-42.

Bülow-Schramm, Margret / Edith Braun (2013): Einleitung, in: Edith Braun / André Donk / Margret Bülow-Schramm (Hrsg.), AHELO Goes Germany? Dokumentation des GfHf- & HIS-HF-Workshops, HIS: Forum Hochschule 2 / 2013, Hannover, S. 1-4.

Carstensen, Claus H. (2006): Technische Grundlagen für die Messwiederholung, in: Manfred Prenzel / Jürgen Baumert / Werner Blum / Rainer Lehmann / Detlev Leutner / Michael Neubrand / Reinhard Pekrun / Jürgen Rost / Ulrich Schiefele (Hrsg.), PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres, Münster, S. 309-323.

de Ayala, Rafael J. (2009): The Theory and Practice of Item Response Theory, New York / NY.

Deutsche Gesellschaft für Soziologie (DGS) (2005): Empfehlungen der Deutschen Gesellschaft für Soziologie (DGS) zur Ausgestaltung soziologischer Bachelor- und Master-Studiengänge, abrufbar unter: www.soziologie.de/uploads/media/BA-MA-Studienempfehlungen-DRUCKF-051212.pdf, letztes Abrufdatum: 17.7.2013.

Deutsche Gesellschaft für Soziologie (DGS) (2012): Wissenschaftliche Evaluation ja – CHE-Ranking nein. Methodische Probleme und politische Implikationen des CHE-Hochschulrankings, abrufbar unter: www.soziologie.de/che, letztes Abrufdatum: 17.7.2013.

Förster, Manuel / Roland Happ / Olga Zlatkin-Troitschanskaia (2012): Valide Erfassung des volkswirtschaftlichen Fachwissens von Studierenden der Wirtschaftswissenschaften und der Wirtschaftspädagogik – eine Untersuchung der diagnostischen Eignung des Wirtschaftskundlichen Bildungstests (WBT), in: bwp@, Berufs- und Wirtschaftspädagogik – online 22, S. 1-21.

Förster, Manuel / Olga Zlatkin-Troitschanskaia (2010): Wirtschaftliche Fachkompetenz bei Studierenden mit und ohne Lehramtsperspektive in den Diplom- und Bachelorstudiengängen – Messverfahren und erste Befunde, in: Klaus Beck / Olga Zlatkin-Troitschanskaia (Hrsg.), Lehrerprofessionalität – Was wir wissen und was wir wissen müssen (Lehrerbildung auf dem Prüfstand, Sonderheft), Landau, S. 106-125.

Frosch, Ulrike (2012): Pädagogische Diagnostik im Spiegel klassischer Lerntheorien. Aktuelle Herausforderungen im Kompetenzdiskurs angesichts einer „Theorie-Methoden-Passung“, in: bwp@, Berufs- und Wirtschaftspädagogik – online 22, S. 1-12.

Geiser, Christian / Michael Eid (2010): Item-Response-Theorie, in: Christof Wolf / Henning Best (Hrsg.), Handbuch der sozialwissenschaftlichen Datenanalyse, Wiesbaden, S. 311-332.

Hartig, Johannes / Eckhard Klieme (2006): Kompetenz und Kompetenzdiagnostik, in: Karl Schweizer (Hrsg.), Leistung und Leistungsdiagnostik, Heidelberg, S. 127-143.

Hadjar, Andreas (Hrsg.) (2011): Geschlechtsspezifische Bildungsungleichheiten, Wiesbaden.

Ho Yu, Chong (2012): A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling (Working Paper), abrufbar unter: http://www.creative-wisdom.com, letztes Abrufdatum: 6.3.2014.

Jerusalem, Matthias / Schwarzer, Ralf (2012): Dimensionen der Selbstwirksamkeit, in: Angelika Glöckner-Rist (Hrsg.), Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 15.00, Bonn: GESIS.

Jonkisz, Ewa / Helfried Moosbrugger / Holger Brandt (2012): Planung und Entwicklung von Tests und Fragebogen, in: Helfried Moosbrugger / Augustin Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion, 2. Auflage, Berlin – Heidelberg, S. 27-74.

Kleickmann, Thilo / Dirk Richter / Mareike Kunter / Jürgen Elsner / Michael Besser / Stefan Krauss / Jürgen Baumert (2013): Teachers' Content Knowledge and Pedagogical Content Knowledge: The Role of Structural Differences in Teacher Education, in: Journal of Teacher Education 64, S. 90-106.

Klieme, Eckhard / Johannes Hartig (2007): Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs, in: Manfred Prenzel / Ingrid Gogolin / Heinz-Hermann Krüger (Hrsg.), Kompetenzdiagnostik, Wiesbaden, S. 11-29.

Klieme, Eckhard / Detlev Leutner (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG, in: Zeitschrift für Pädagogik 52, S. 876-903.

Krathwohl, David R. (2002): A Revision of Bloom's Taxonomy: An Overview, in: Theory into Practice 41, S. 212-218.

Kubinger, Klaus D. / Clemens Draxler (2007): Probleme bei der Testkonstruktion nach dem Rasch-Modell, in: Diagnostica 53, S. 131-143.

Mair, Patrick / Reinhold Hatzinger / Marco J. Maier (2013): Package ‚eRm' (Version 0.15-1), abrufbar unter: www.cran.r-project.org/web/packages/eRm/eRm.pdf, letztes Abrufdatum: 6.3.2014.

Masters, Geoff (1982): A Rasch Model for Partial Credit Scoring, in: Psychometrika 47, S. 149-174.

Moosbrugger, Helfried (2012): Item-Response-Theorie (IRT), in: Helfried Moosbrugger / Augustin Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion, 2. Auflage, Berlin – Heidelberg, S. 227-274.

Müller-Benedict, Volker / Elena Tsarouha (2011): Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen, in: Zeitschrift für Soziologie 40, S. 388-409.

Organisation for Economic Cooperation and Development (OECD) (Hrsg.) (2013): Assessment of Higher Education Learning Outcomes (AHELO). Feasibility Study Report Volume 2 – Data Analysis and National Experiences: OECD.

Pant, Hans Anand / Katrin Böhme / Olaf Köller (2012): Das Kompetenzkonzept der Bildungsstandards und die Entwicklung von Kompetenzstufenmodellen, in: Petra Stanat / Hans Anand Pant / Katrin Böhme / Dirk Richter (Hrsg.), Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011, Münster, S. 49-55.

Prenzel, Manfred / Cordula Artelt / Jürgen Baumert / Werner Blum / Marcus Hammann / Eckhard Klieme / Reinhard Pekrun (PISA-Konsortium Deutschland) (Hrsg.) (2008), PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich, Münster.

Prenzel, Manfred / Ingrid Gogolin / Heinz-Hermann Krüger (Hrsg.) (2008): Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft, Sonderheft 8 / 2007, Wiesbaden.

Rasch, Georg (1960): Probabilistic Models for some Intelligence and Attainment Tests, Kopenhagen: The Danish Institute for Educational Research.

Riese, Josef / Peter Reinhold (2012): Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. Empirische Hinweise für eine Verbesserung des Lehramtsstudiums, in: Zeitschrift für Erziehungswissenschaft 15, S. 111-143.

Rizopoulos, Dimitris (2012): Package ‚ltm'. Latent Trait Models under IRT (Version 0.9-9), abrufbar unter: www.cran.r-project.org/web/packages/ltm/ltm.pdf, letztes Abrufdatum: 6.3.2014.

Rost, Jürgen (2004): Lehrbuch Testtheorie – Testkonstruktion, 2. Auflage, Bern.

Roth, Heinrich (1971): Pädagogische Anthropologie. Band 2: Entwicklung und Erziehung. Grundlagen einer Entwicklungspädagogik, Hannover.

Samejima, Fumiko (1969): Estimation of Latent Ability Using a Response Pattern of Graded Scores, in: Psychometrika Monograph Supplement 17.

Spiel, Christiane / Barbara Schober / Margarete Litzenberger (2008): Projektbericht: Evaluation der Eignungstests für das Medizinstudium in Österreich, Wien: Universität Wien (Evaluationsprojekt im Auftrag des Bundesministeriums für Wissenschaft und Forschung).

Stanat, Petra / Hans Anand Pant / Katrin Böhme / Dirk Richter (Hrsg.) (2012): Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächer Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011, Münster.

StataCorp (Hrsg.) (2011): Multiple Imputation Reference Manual Release 12, College Station / TX.

Strobl, Carolin (2010): Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis, München – Mering.

Strobl, Carolin (2012): Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis, 2. Auflage, München – Mering.

Strobl, Carolin / Julia Kopf (2010): Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben, in: Sabine Trepte / Markus Verbeet (Hrsg.), Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test, Wiesbaden, S. 255-272.

Tatto, Maria Teresa / Ray Peck / John Schwille / Kiril Bankov / Sharon L. Senk / Michael Rodriguez / Lawrence Ingvarson / Mark Reckase / Glenn Rowley (2012): Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries. Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M), Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Tremblay, Karine / Diane Lalancette / Deborah Roseveare (2012): Assessment of Higher Education Learning Outcomes (AHELO), Feasibility Study Report Volume 1 – Design and Implementation: OECD.

Trepte, Sabine / Markus Verbeet (Hrsg.) (2010): Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test, Wiesbaden.

Van der Linden, Wim J. / Ronald K. Hambleton (Hrsg.) (1997): Handbook of Modern Item Response Theory, New York / NY.

Walstad, William B. / Denise Robson (1997): Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics, in: The Journal of Economic Education 28, S. 155-171.

Weinert, Franz E. (2001): Concept of Competence: A Conceptual Clarification, in: Dominique Simone Rychen / Laura Hersh Salganik (Hrsg.), Defining and Selecting Key Competencies, Seattle / WA, S. 45-65.

Wilson, Mark (2005): Constructiong Measures. An Item Response Modeling Approach, Mahwah / NJ.

Winkelmann, Henrik / Alexander Robitzsch (2009): Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität, in: Dietlinde Granzer / Olaf Köller / Albert Bremerich-Vos (Hrsg.), Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule, Weinheim – Basel, S. 169-196.

Wissenschaftsrat (2012): Prüfungsnoten an Hochschulen im Prüfungsjahr 2010. Arbeitsbericht mit einem Wissenschaftspolitischen Kommentar des Wissenschaftsrates, Hamburg.

Wu, Margaret (2010): Comparing the Similarities and Differences of PISA 2003 and TIMSS (OECD Education Working Papers No. 32): OECD Publishing.

Zlatkin-Troitschanskaia, Olga / Jana Seidel (2011): Kompetenz und ihre Erfassung – das neue „Theorie-Empirie-Problem“ der empirischen Bildungsforschung?, in: Olga Zlatkin-Troitschanskaia (Hrsg.), Stationen Empirischer Bildungsforschung. Traditionslinien und Perspektiven, Wiesbaden, S. 218-233.

Zlatkin-Troitschanskaia, Olga / Sigrid Blömeke / Christiane Kuhn / Christiane Buchholtz (2012): Wirksamkeitsprüfungen im Hochschulbereich – Aufgaben und Herausforderungen des BMBF-Forschungsprogramms „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“, in: Zeitschrift für Evaluation 11, S. 95-103.

Zlatkin-Troitschanskaia, Olga / Manuel Förster / Roland Happ (2012): Bologna-Reform – Ergebnisse aus einer vergleichenden empirischen Studie zwischen den auslaufenden Diplom- und den neuen Bachelor-/Masterstudiengängen, in: Zeitschrift für Berufs- und Wirtschaftspädagogik 108, S. 420-437.

Dr. Felix Wolter
Dr. Jürgen Schiener
Johannes Gutenberg-Universität Mainz
Institut für Soziologie
Jakob-Welder-Weg 12
55128 Mainz
felix.wolter@uni-mainz.de
juergen.schiener@uni-mainz.de