QUALITEACH
TEACHING TALENT CENTER

Nastassja Simeth und Ernst Hany

Entwicklung eines Performance Assessment-Verfahrens mit Multiplen Mini-Interviews für Lehramtsstudierende

Statusbericht 2017 aus dem „Teaching Talent Center“

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Performance Assessment im Teaching Talent Center

■ Aufgabenstellungen des Teaching Talent Centers

Im Rahmen der gemeinsamen Qualitätsoffensive Lehrerbildung von Bund und Ländern erhält die Universität Erfurt in diesen Jahren Fördermittel des Bundesministeriums für Bildung und Forschung für die Fort- und Weiterentwicklung der Lehrerbildung. Eines der Teilprojekte des Erfurter Vorhabens QUALITEACH ist das „Teaching Talent Center", das es sich zur Aufgabe gemacht hat, Talente für den Lehrberuf zu identifizieren und zu fördern sowie den Lehrberuf insgesamt für talentierte junge Menschen attraktiv zu machen. Eine wichtige Basis für diese Arbeit ist die Verfügbarkeit von Messinstrumenten, mit denen Persönlichkeitsmerkmale von Lehramtsstudierenden oder gegebenenfalls von Bewerberinnen und Bewerbern für das Lehramtsstudium erfasst werden können. Der Einsatz dieser Instrumente ist derzeit nicht im Rahmen eines Zulassungsverfahrens zum Lehramtsstudium mit dem Ziel der Selektion vorgesehen. Vielmehr soll den Studierenden die Möglichkeit verschafft werden, die eigenen Stärken und Schwächen mit Bezug zum Lehrberuf frühzeitig zutreffend einschätzen zu können und nachfolgend eigenverantwortlich Entscheidungen zu treffen. Diese Entscheidungen können sich auf die Wahl des Studiums und des Berufs beziehen, sie können aber genauso darauf abzielen, sich selbst noch besser auf die Anforderungen des Lehrberufs vorzubereiten. Zu diesem Zweck will das Teaching Talent Center Beratungs- und Trainingsangebote besonders in Bezug auf Schlüsselkompetenzen entwickeln.

■ Ziel dieses Berichts

Ziel dieses Berichts ist, den Stand der Entwicklungsarbeiten im Hinblick auf verhaltensdiagnostische Verfahren zum Berichtszeitpunkt Mitte 2017 vorzustellen und so über die geleistete Arbeit zu informieren. Die gewonnenen Erkenntnisse werden dargestellt und die ermittelten Probleme reflektiert. Ferner werden die in der kommenden Zeit geplanten Arbeitsschritte vorgestellt und die Arbeitsziele gegebenenfalls im Licht der gewonnenen Erkenntnisse angepasst.

Zitiervorschlag für dieses Papier:

Simeth, N. & Hany, E. (2017). *Entwicklung eines Performance Assessment-Verfahrens mit Multiplen Mini-Interviews für Lehramtsstudierende. Statusbericht 2017 aus dem „Teaching Talent Center".* Erfurt: Universität Erfurt, Erfurt School of Education, Projekt QUALITEACH.

Kontaktadresse: Nastassja Simeth, M.Sc., Fachgebiet Psychologie, Universität Erfurt, Nordhäuser Str. 63, D-99089 Erfurt. E-Mail: nastassja.simeth@uni-erfurt.de

Inhaltsverzeichnis

Ziele, Inhalte und methodische Ansätze der Eignungsdiagnostik für das Lehramt

Auf der Grundlage der subjektiven Evidenz von Einzelfällen neigen viele Menschen zu der Annahme, dass etliche Personen den Lehrberuf ergriffen haben, obwohl sie sich dafür nicht wirklich eignen. So liegt die Vorstellung nahe, ein Verfahren zu entwickeln, das garantieren möge, dass an unseren Schulen nur solche Personen tätig sind, die den vielfältigen und durchaus hochgesteckten Ansprüchen der Betrachter*innen entsprechen. Am besten, so stellen sich das wohl viele vor, sollte man die Eignung für den Lehrberuf bereits vor Aufnahme des entsprechenden Studiums erkennen und ungeeigneten Personen entweder dringend von der Aufnahme des lehramtsbezogenen Studiums abraten oder sie mit einem standardisierten Auswahlverfahren vom Studium ausschließen.

Nun stellt die Eignungsdiagnostik bereits seit etwa 100 Jahren ein wissenschaftlich fundiertes Instrumentarium für die Auswahl von Bewerber*innen in der Arbeitswelt bereit (Moede, 1930). Es kann daher nicht an den theoretischen Grundlagen oder den verfügbaren Messverfahren liegen, dass es bisher nicht gelungen ist, die oben geschilderten Vorstellungen in einem Routineverfahren abzubilden.

Bei genauerem Hinsehen werden mehrere grundlegende Probleme offenbar: So ändern sich die Anforderungen an den Lehrberuf fortlaufend und es ist schlechterdings unmöglich, eine Eignungsdiagnostik für Anforderungen vorzunehmen, die vielleicht erst zehn oder 20 Jahre später gültig sind (Nolle, 2016). Ferner ist es zentrale Aufgabe der Hochschulen, den Studierenden diejenigen Kompetenzen zu vermitteln, die sie für die erfolgreiche Ausübung des Lehrberufs benötigen. Dabei ist weiterhin unklar, in welchem Umfang stabile Persönlichkeitsmerkmale oder habituelle Strategien der Stressbewältigung und Arbeitsorganisation den Verlauf des Studiums und die Qualität der Berufsausübung prägen. Rothland (2013, S. 88) erteilt entsprechenden Hoffnungen beim heutigen Stand der Forschung eine klare Absage: „Vielmehr ist es grundsätzlich bislang nicht ausgemacht, geschweige denn empirisch längsschnittlich abgesichert, wie die Eignung für den Lehrerberuf überhaupt valide erfasst und langfristig prognostiziert werden kann bzw. wie der Berufserfolg als Kriterium erschöpfend und umfassend zu operationalisieren ist.“

Die Bedenken des Autors werden von vielen Akteuren geteilt. Zudem können Rothland und Terhart (2011, S. 637) feststellen, dass es bislang niemand gewagt hat, ein *berufsspezifisches* selektives Auswahlverfahren jenseits *allgemeiner* Eignungskriterien, wie sie sich beispielsweise in der Abiturnote niederschlagen, für den Lehrberuf einzuführen: „Die diagnostischen Schwierigkeiten in Verbindung mit der rechtlichen Problematik haben dazu geführt, dass bislang in Deutschland noch keine Universität ein tatsächlich fremd-selektives berufsspezifisches Eignungsfeststellungsverfahren beim Zugang zu Lehramtsstudiengängen einsetzt“. Vielmehr setzen die Hochschulen in der Regel auf Selbstselektion und den Ausschluss Ungeeigneter durch das Nichtbestehen der akademischen Prüfungen.

Dennoch setzen zahlreiche Hochschulen im In- und Ausland auf aktive Maßnahmen zur Steuerung der Berufswahl. Häufig werden Studieninteressierte, die den Lehrberuf anstreben, darauf hingewiesen oder sogar dazu verpflichtet, Selbsterkundungsverfahren durchzuführen und die eigene Eignung für den Lehrberuf gründlich zu bedenken. Diese Verfahren bieten nicht nur die Erfassung von beruflichen Interessen und lehrberufsspezifischen Persönlichkeitszügen an, sondern sie informieren vor allem umfassend über die Herausforderungen

des Lehrberufs und schließen damit Wissenslücken auf Seiten der Interessierten. Solche Maßnahmen zur Förderung der Auseinandersetzung mit der eigenen Berufswahl können auch nach Aufnahme des Studiums vorgesehen sein, vor allem dann, wenn es gilt, die ersten Praktika zu absolvieren, dabei die Schülerperspektive zu verlassen und Schule aus einer Lehrendenperspektive wahrzunehmen.

Durchgeführte diagnostische Maßnahmen erfüllen dabei weniger eine Selektionsfunktion, sondern vielmehr eine Modifikationsfunktion (nach Krapp, 1979), indem sie Daten liefern, die als Ausgangspunkt für Beratungs- und Trainingsmaßnahmen dienen können. Dabei ist es nachrangig, ob sich die diagnostischen Erhebungen auf allgemeine Persönlichkeitsmerkmale, die für das zielstrebige und erfolgreiche Studieren erforderlich sind, oder auf spezifische Dispositionen und habituelle Merkmale, die für den Lehrberuf relevant sind, beziehen. Denn ob problematische Ausprägungen dieser Merkmale zu persönlichen Entscheidungen hinsichtlich der Studien- und Berufslaufbahn oder hinsichtlich der Wahrnehmung von Beratungs- und Trainingsangeboten führt, hängt allein von der Problemwahrnehmung der Studierenden ab. Diese Assessmentverfahren können daher vor allem als Servicefunktion für die Lehramtsstudierenden gesehen werden, die ihnen helfen, die eigene Entwicklung im Sinne eines erfolgreichen „life design“ voranzubringen. Die Validität dieser Messverfahren bestimmt sich also zu einem wesentlichen Teil im Sinne von Messick (1995) aus den Konsequenzen, die die Beteiligten daraus ziehen.

Wenn es Studierenden hilft, etwas über ihre beruflichen Interessen und der Passung zu Berufsfeldern im Schulwesen zu erfahren, weil sie daraus Entscheidungen ableiten, so sind diese Messungen valide. Genauso valide ist aber auch die Erfassung der psychischen Labilität, der Strategien zur Stressbewältigung, des Zeit- und Aufgabenmanagements, der sozialen Kompetenz, der persönlichen Konfliktverarbeitung oder der ressourcenbewussten Lebensführung – und zwar immer dann, wenn Studierende erkennen, dass sie in diesen Bereichen Entwicklungs- und Beratungsbedarf haben. Im Unterschied zu ergebnisorientierten Eignungsmerkmalen spricht Nolle (2016) hier von „entwicklungsprozessorientierten Eignungsmerkmalen“ und meint damit Personenmerkmale, deren Ausprägungen entscheidend dafür sind, ob Studierende das gewählte Studium für den Kompetenzerwerb und die Persönlichkeitsentwicklung überhaupt nutzen können. Die Studien von Košinár (2014) oder Gottein (2016) zeigen, wie mühsam oder manchmal unmöglich es ist, subjektive pädagogische Theorien im Laufe des Studiums durch Information, Erfahrung oder angeleitete Reflexion zu modifizieren. So erweisen sich ungünstige Wissensstrukturen als ein weiterer Ansatzpunkt für die Eignungsdiagnostik. Ein passendes Eignungsfeststellungsverfahren könnte helfen, späterer „pädagogischer Borniertheit“ einen Riegel vorzuschieben.

Wir stellen fest, dass es nicht nur die Kompetenzmodelle für die *erfolgreiche Ausübung* des Lehrberufs sind, aus denen diagnostisch nutzbare Konstrukte ableitbar sind. Modelle des *Kompetenzerwerbs* im Rahmen einer selbstverantwortlichen akademischen Ausbildung, Modelle der *Persönlichkeitsentwicklung* im frühen Erwachsenenalter und Modelle der Wirksamkeit und Veränderbarkeit *naiven Wissens* können genauso Ansatzpunkte liefern, um Assessmentsysteme zu generieren, die den Studierenden wichtige Entwicklungsimpulse vermitteln können.

In welcher Form dazu passende Messverfahren erstellt oder ausgewählt werden, muss je nach theoretischem Konstrukt entschieden werden. So liegen für die Erfassung beruflicher Interessen sehr brauchbare Papier- und Bleistift-Verfahren vor. Für die Erfassung sozialer Kompetenzen u. dgl. werden eher Verhaltensbeobachtungen und Verhaltensbeurteilungen

herangezogen, während subjektive pädagogische Konzepte vielleicht am besten in Form eines Interviews ermittelt werden können. Grundsätzlich müssen bei der Wahl jedes Messverfahrens die etablierten Testgütekriterien im Vordergrund stehen. Hier sind die sogenannten Nebengütekriterien, allen voran Ökonomie und Fairness, nicht zu vernachlässigen, da perspektivisch eine große Zahl von Studierenden unterschiedlicher Herkunft und mit unterschiedlichen Biografien in die Assessment- und Beratungsverfahren einbezogen werden sollen. Der dabei erforderliche Aufwand und die möglichen Benachteiligungen bestimmter Personengruppen sind sehr genau zu prüfen, bevor solch ein Verfahren routinemäßig etabliert werden kann.

Performance Assessment als Methode der Erfassung von Handlungskompetenzen

Im Bereich der schulischen Kompetenzmessung wurden in den letzten Jahren große Anstrengungen unternommen, um das, was Schüler*innen im schulischen Unterricht gelernt haben, in möglichst objektiver Form, d. h. mit geschlossenen Aufgaben, ökonomisch zu erfassen. In den groß angelegten internationalen Erhebungen wäre es nicht möglich gewesen, das Wissen und Können der Schüler*innen mit offenen Aufgaben zu erfassen, die im Nachhinein einen hohen Auswertungs- und Interpretationsaufwand erfordern. Dennoch gibt es zahlreiche Könnensbereiche, in denen es unumstritten ist, dass zum Nachweis genuine *Verhaltensleistungen* erforderlich sind. Exemplarisch sei die praktische Führerscheinprüfung genannt, in der Anwärter*innen zeigen müssen, dass sie die Führung des Kraftfahrzeugs beherrschen. Eine rein theoretische Prüfung ist für den Nachweis der geforderten Kompetenzen offenbar nicht ausreichend. Auch im akademischen Bereich kennen wir verhaltensorientierte Prüfungen beim Eignungsnachweis für sportliche, künstlerische oder musische Studiengänge.

Der Fachausdruck für Messungen, die von den Probanden ein konkretes Handeln zur Bewältigung von Leistungsansprüchen erfordern, heißt „Performance Assessment“. Nach Palm (2001) sind dafür drei Merkmale kennzeichnend:

- Die *Aufgaben* erfordern selbst gestaltetes Handeln, Auftreten, Produzieren oder Präsentieren.
- Die *Situationen* (Aufgabenkontexte) ähneln dem wahren Leben bzw. authentischen Problemstellungen (high fidelity simulations) im beruflichen Kontext.
- Die gezeigten *Leistungen* entsprechen relativ stark den im Berufseinsatz geforderten.

Entscheidend für den Einsatz eines Performance Assessment ist also die hohe Ähnlichkeit zwischen den in den standardisierten Messungen und den in der vorgesehenen Praxis geforderten Leistungen. Deshalb wird die praktische Führerscheinprüfung auch nicht an einem Fahrsimulator, sondern mit einem handelsüblichen Fahrzeug und im authentischen Straßenverkehr durchgeführt. Um die „high fidelity“ der Prüfungssituation zu gewährleisten, müssen (nach Palm, 2001) mehrere Forderungen erfüllt sein:

- Gefordert werden im Rahmen der für diagnostische Zwecke zu erbringenden Leistung dieselben Denkprozesse, Handlungen, Produkte usw. wie in der Praxis;

- die Aufgabenbedingungen sollten weitgehend der Realität (z. B. Zeitdruck, Hilfsmittel, Team) entsprechen;
- die Aufgabeneinkleidung sollte sich an der Realität orientieren (z. B. „Stellen Sie sich vor, Sie sind Lehrerin an einer Gemeinschaftsschule …“);
- die Bewertungskriterien sollen der Praxis entstammen und möglichst mit Personen aus der Praxis entwickelt worden sein.

Um diese Kriterien zu erfüllen, wäre es am einfachsten, die Kandidatinnen und Kandidaten würden probehalber die real vorgesehene Leistung erbringen. Dies ist überall dort der Fall, wo sich mit Bewerberinnen und Bewerbern auf eine Probezeit oder eine enger begrenzte Arbeitsprobe geeinigt wird, in der die am angestrebten Arbeitsplatz üblichen Tätigkeiten bereits authentisch durchgeführt werden. Bekannt ist etwa aus früheren Zeiten, dass während eines Bewerbungsgesprächs für einen Sekretariatsposten ein DIN-konformer Brief nach Diktat möglichst fehlerfrei zu erstellen war. Bei vielen Tätigkeiten wäre es aber zu aufwändig und zu riskant, Personen einzusetzen, deren Wissen und Können durch den Einsatz erst geprüft werden sollen. Man denke etwa an den Einsatz von Flugsimulatoren in der Pilotenausbildung oder an schulische Unterrichtsstunden im Rahmen der universitären Lehrerbildung, die nach gründlicher Vorbereitung unter Aufsicht durchgeführt werden. Es ist in vielen Fällen zur Vermeidung kritischer Konsequenzen sinnvoll, künstliche Handlungssituationen herzustellen, die auf die wesentlichen Leistungsanforderungen fokussieren und die von Beobachter*innen konzentriert verfolgt werden können.

Stiggins (1987) macht deutlich, dass bei der Gestaltung eines Performance Assessment mehrere Festlegungen im Sinne von überlegten Entscheidungen zu treffen sind. Zunächst ist die Fertigkeit oder Kompetenz festzulegen, über deren Ausprägung die vorgesehene Messung eine Aussage treffen soll. Als nächstes ist eine Aufgabe zu formulieren, bei deren Bearbeitung eine Darbietung gezeigt werden muss, die als repräsentativ für die zu erfassende Kompetenz gelten kann. Für diese Darbietung gilt es Leistungskriterien festzulegen, nach denen die Performance beurteilt wird. Im einfachsten Fall ist das Kriterium die Zeit, die für die Erledigung der Aufgabe benötigt wird, wie dies beispielsweise bei Lauf- oder Skisportarten Usus ist. In komplexeren Fällen (beispielsweise beim Kunstturnen oder beim Turniertanz, um im Sportbereich zu bleiben) sind solche Beurteilungskriterien mehrdimensional. Stiggins weist darauf hin, dass auch zu entscheiden ist, ob die zu erfassende Kompetenz im Rahmen von Alltagshandlungen oder im Rahmen definierter Erfassungssituationen beobachtet werden soll. Ferner gilt es zu entscheiden, ob die Beobachtung verdeckt oder mit Wissen aller Beteiligten durchgeführt wird und ob eine ein- oder mehrmalige Beobachtung Grundlage der Kompetenzeinschätzung sein soll.

Formen des Performance Assessment finden sich auch in Ausbildungsgängen für den Lehrberuf. Das Beispiel par excellence ist die Lehrprobe, in der Studienreferendar*innen eine vollständige Unterrichtsstunde planen, vorbereiten, durchführen und anschließend analysieren. Nun ist nachvollziehbar, dass eine so komplexe Ausbildung wie die erste bzw. die zweite Phase der Lehrerbildung mit einer möglichst authentischen Leistungsmessung abschließt. Wie lassen sich dagegen Formen des Performance Assessment zu Beginn der Ausbildung, im Sinne einer eignungsdiagnostischen Untersuchung rechtfertigen? Die erforderlichen Kompetenzen für den Lehrberuf sind zu diesem Zeitpunkt noch nicht entwickelt, da das Studium noch keine Wirkung entfalten konnte.

Die Begründung für den Einsatz von Verhaltensproben findet sich letztlich im Kompetenzbegriff, wie er die heutige Diskussion um Bildungsprozesse und Bildungsprogramme prägt. Nach diesem Paradigma (Klieme et al., 2003) sind für die Bewältigung authentischer Problemsituationen nicht nur einzelne Persönlichkeitsmerkmale, Fertigkeiten oder Wissensbestände erforderlich, vielmehr ist es das Zusammenspiel der passenden Komponenten und ihre strategische Nutzung bei der Bewältigung des Problems.

So weisen beispielsweise die Konzepte „soziale Kompetenz“ oder „emotionale Kompetenz“ bereits durch ihre Begrifflichkeit darauf hin, dass es sich jeweils um das Zusammenspiel mehrerer Komponenten handelt, die für die Bewältigung einer sozial bzw. emotional anspruchsvollen Situation aktiviert und koordiniert werden müssen. Gerade bei Anforderungen in sozialen Situationen und bei der argumentativen Darlegung von Positionen muss davon ausgegangen werden, dass Persönlichkeitsmerkmale, Präsentationsstrategien, logisches Denken, Kompromissbereitschaft, Flexibilität usw. zusammenwirken. Außerdem ist anzunehmen, das Argumentation, Präsentation, Kommunikation, Konfliktlösung sowie genuin pädagogische Tätigkeiten der Anleitung, der erzieherischen Einwirkung, der Arbeitsteilung mit Kolleg*innen usw. in der Tat bedeutsam für die Bewältigung regelmäßig wiederkehrender Anforderungen im Lehrberuf sind. Ferner wird erwartet, dass die individuelle Bewältigung dieser Anforderungen zumindest teilweise von relativ stabilen Persönlichkeitszügen geprägt ist, deren Ausformung bereits zu Beginn des Studiums erfassbar ist.

Das Performance Assessment, mit dem Verhaltensleistungen vor oder zu Beginn des Studiums erfasst werden, entspricht damit im Grunde einer Verhaltensbeobachtung bzw. -beurteilung von relativ stabilen Persönlichkeitszügen, insoweit diese für die Bewältigung berufsnaher Anforderungen eingesetzt werden. Bei stark ungünstig ausgeprägten und das Sozialverhalten negativ beeinflussenden Persönlichkeitsdispositionen wird Studieninteressierten am besten vom Lehrberuf abgeraten; bei schwächeren Ausprägungen können hingegen Maßnahmen ergriffen werden, um durch Selbsterkenntnis, Selbstreflexion und gezieltes Üben kompensatorische Strategien zu entwickeln, die die erfolgreiche Bewältigung der berufsrelevanten Anforderungssituationen ermöglichen.

Nach den Ausführungen von Stemmler und Margraf-Stiksrud (2015) sind Verhaltensbeobachtungen von Persönlichkeitszügen nur dann sinnvoll, wenn das gezeigte Verhalten von *kontextualisierten Dispositionen* geprägt ist, die über gewisse Zeiträume und gewisse Situationsklassen intraindividuell stabil und interindividuell unterschiedlich ausgeprägt sind. Es müssen also zumindest Verhaltensgewohnheiten vorliegen, damit es sich lohnt, diese auf dem Weg der direkten Verhaltensbeobachtung zu erfassen und die Ergebnisse für weiterführende Entscheidungen zu nutzen.

Die Autoren weisen darauf hin, dass mehrere Bedingungen gegeben sein müssen, damit Verhaltensdispositionen im konkreten Verhalten erfasst werden können:

- So muss es berufsnahe, aber im Labor herstellbare Situationen geben, in denen die kontextualisierte Disposition aktiviert wird, d. h. zu beobachtbarem Verhalten führt. Die Qualität einer Situation, die ausgewählte Disposition zum Vorschein zu bringen, wird als *Situationsvalenz* bezeichnet. Mit Situationsdefizienz wird hingegen das Ausmaß einer Situation bezeichnet, andere als die eigentlich erwünschten Dispositionen zu aktivieren.

- Ferner muss es möglich sein, bestimmte Verhaltensweisen der ausgewählten Disposition eindeutig zuzuordnen. Ebenso müssen diese Verhaltensweisen mehrfach in der gestalteten Situation auftreten.

- Die relevanten Verhaltensweisen müssen von Beobachtenden eindeutig wahrnehmbar und es muss ein Aufzeichnungsinstrument verfügbar sein, mit dem die Beobachtungen als solche – oder verarbeitet zu Beurteilungen – registriert werden können.

Stemmler und Margraf-Stiksrud (2015) beschreiben *am Beispiel der sozialen Kompetenz* die notwendigen Schritte bei der Konstruktion eines Performance Assessment. Dieser Ablauf wurde im Frühjahr 2017 denjenigen Personen nahegelegt, die sich mit der Entwicklung und Erprobung von geeigneten Anforderungssituationen beschäftigten (siehe das Kapitel „Laufende Arbeiten“ ab S. 27).

- Zunächst ist das zentrale Konzept festzulegen, in diesem Fall „soziale Fertigkeiten“ als Konkretisierung des übergeordneten Konstrukts „Soziale Kompetenz“.

- Anschließend ist das Konstrukt in handhabbare Teilkonstrukte zu unterteilen. In Frage kommen beispielsweise Gesprächsführung, Teamfähigkeit, Auftreten, Einfühlungsvermögen, Kontaktfähigkeit, Überzeugungskraft, Verhandlungsgeschick, Konfliktverhalten oder Integrationsfähigkeit.

- Zur Erfassung dieser spezifischen Fertigkeiten sind geeignete Situationen (d. h. kontextualisierte Aufgaben) und Beobachtungsmerkmale, d. h. die relevanten Verhaltensweisen, die für das Vorliegen der jeweiligen Fertigkeit sprechen, zu entwickeln bzw. festzulegen.

- Für die Erfassung der Verhaltensweisen und die Verrechnung zu Merkmalsausprägungen sind Beobachtungsdesigns, Registrierungsverfahren und Rechenalgorithmen festzulegen.

- Gegebenenfalls ist festzulegen, wie die im vorhergehenden Schritt ermittelten Indikatoren zu einer individuellen Gesamtausprägung für das Konstrukt „soziale Kompetenz“ weiter verrechnet werden.

Jede Art des Performance Assessment erfordert eine Erfassung der gezeigten Leistung in Form der Registrierung von Beobachtungen oder der Bestimmung von Beurteilungen. In der Regel sind an diesen Verarbeitungsschritten Personen beteiligt, während die Auswertung von geschlossenen Aufgaben (wie sie beispielsweise bei Leistungstests zum Einsatz kommen) auch computergestützt erfolgen kann. Das Performance Assessment ist daher nicht nur besonders aufwändig, sondern durch die Beteiligung verschiedener Personen bei der Durchführung und Auswertung auch besonders fehleranfällig. Wie Studien zur testtheoretischen Qualität von Auswahlgesprächen im Lehramt zeigen, sind Durchführungs- und Auswertungsobjektivität nur schwer zu sichern (Frost, 2015, S. 272). Unterrichtsbeurteilungen, selbst wenn sie durch Fachkolleg*innen oder Mitglieder der Schulleitung durchgeführt werden, fallen unzuverlässig aus (Ho & Kane, 2013, S. 15). Auch in der eignungsdiagnostischen Nutzung von umfassenden Beobachtungen im Rahmen eines „Assessment Centers“ hat sich starke Ernüchterung breit gemacht, da die Beurteilungen offenbar weder ausreichende Reliabilität noch überzeugende Validität besitzen (Schuler, 2007). Bei der Erfassung von Verhaltensleis-

tungen dürfen folglich keine Wunder, sondern müssen im Gegenteil besonders hohe methodische Herausforderungen und zahlreiche Probleme bei der Qualitätssicherung erwartet werden.

Das Multiple Mini-Interview (MMI)

Das multiple Mini-Interview (MMI) ist ein qualitatives Instrument der Personalauswahl, welches als Alternative zu unstrukturierten Interviews an der McMaster-Universität in Hamilton (Kanada) zur Auswahl von Studierenden für das Medizinstudium entwickelt wurde. Insgesamt stellt ein MMI die Summe von mehreren kleinen Aufgabeneinheiten dar, die ähnlich denen eines Assessment Centers sind. Die Dauer eines solchen MMIs entspricht dagegen einem ausführlichen Interview. Ziel des Verfahrens ist es, wie in der Personalauswahl üblich, diejenigen Bewerbenden zu identifizieren, die sowohl im Studium als auch im späteren Beruf erfolgreich sind. Der Vorteil dieses Auswahlinstrumentes liegt in einer hochgradig standardisierten Struktur, die den Rating-Bias von Beobachtenden minimieren und die Chancengleichheit der Bewerbenden erhöhen soll. Somit besteht das Instrument aus mehreren voneinander unabhängigen Performance Assessments. Während das Verfahren in allen bisher publizierten Studien zur Selektion von Bewerbenden eingesetzt wird (Abbildung 1), gehen wir davon aus, dass es auch zur Beratung und Entwicklung wichtiger Kompetenzfacetten eingesetzt werden kann.

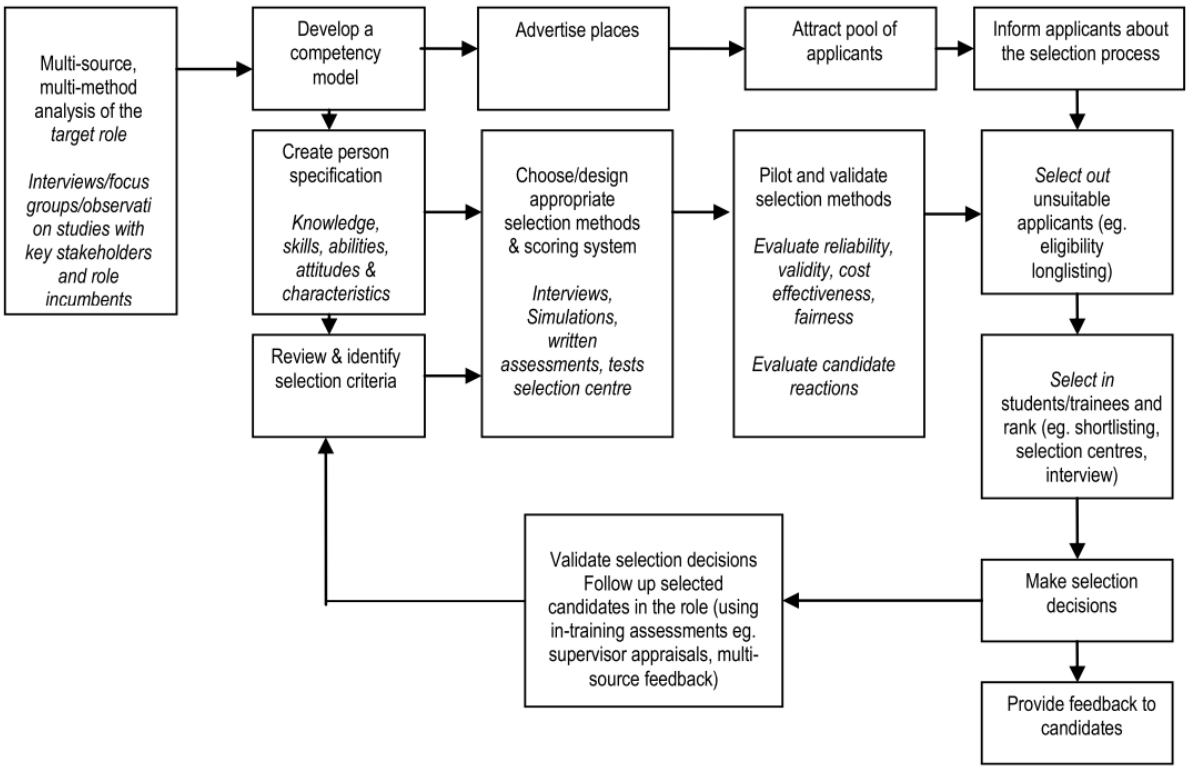

Abbildung 1: Idealer Ablauf bei der Entwicklung eines Selektionsverfahrens (aus Cleland et al., 2012, S. 6)

In Abhebung zu dem in Abbildung 1 gezeigten Verfahren würden wir den Begriff „Selektion“ mit „Inventarisierung“ oder allgemein „Assessment“ ersetzen. Stattdessen fokussieren die laufenden Arbeiten auf veränderliche und entwicklungsfähige Merkmale. Andere Modellkomponenten wie die Aufstellung eines Kompetenzmodells als Basis der Erhebungen, die Bewertung der Reaktion der Testanden und Folgeerhebungen zur weiteren Entwicklung der Studierenden bleiben jedoch erhalten.

■ Aufbau eines MMIs

Das MMI besteht aus mehreren kurzen standardisierten Aufgaben, die von den Bewerbenden nacheinander in einem Rotationsverfahren durchlaufen werden. Daher entspricht die Anzahl der gleichzeitig teilnehmenden Bewerbenden der Anzahl der zu bewältigenden Stationen. Jede dieser Stationen befindet sich in einem separaten Raum und wird von jeweils einem anderen Beobachter-Team betreut. Inhalte einer solchen Station können (situative) Fragen, Rollenspiele oder Diskussionen sein.

Zu Beginn des Verfahrens werden die Bewerbenden zufällig auf die Erhebungs-Stationen verteilt und starten ihr MMI mit eben dieser Station. Der wiederkehrende Ablauf des Verfahrens ist dabei folgendermaßen strukturiert: Mittels eines akustischen Signals werden die Bewerbenden aufgefordert, die Aufgabenstellung ihres Mini-Interviews zu lesen. Diese steht sowohl vor dem Raum als auch später in diesem zur Verfügung. Ein weiteres Signal bedeutet den Bewerbenden das Ende der Vorbereitungszeit und den Beginn der Interview-Situation. Die Bewerbenden dürfen den entsprechenden Raum betreten und die Aufgabenstellung bearbeiten. Ein letztes Signal beendet die Situation und bedeutet den Kandidat*innen, sich zur nächsten Station zu begeben und die neue Aufgabenstellung zu lesen. Diese Prozedur wird wiederholt bis jede*r Teilnehmende alle Interview-Stationen durchlaufen hat. Die Verfahrensdauer pro Person bemisst sich somit aus der Summe der Vorbereitungs- und Performance-Zeit multipliziert mit der Anzahl der zu bearbeitenden Aufgaben.

■ Entwicklungshintergrund

Bester Prädiktor für akademische Leistungen ist eine gute Abiturdurchschnittsnote (Schmidt & Hunter, 1998). Dennoch ist mittlerweile allseits bekannt, dass kognitive Leistungen beruflichen Erfolg nicht alleine vorhersagen können. Daher werden zunehmend auch überfachliche Kompetenzen in Zulassungsentscheidungen miteinbezogen, so auch an der McMaster-Universität, wo die MMIs ihren Ursprung haben. Eine besondere Schwierigkeit ist jedoch die zuverlässige Messung solcher überfachlichen Kompetenzen. Bis einschließlich 2004 wurden diesbezüglich Interviews an der McMaster-Universität durchgeführt. Die Reliabilität dieses Auswahlinstrumentes variiert allerdings stark zwischen $r = .15$ und $r = .95$. Fast 60 % der Varianz können dabei auf Beurteiler-Effekte (z.B. Tendenz zur Milde/Härte, Halo-Effekt) zurückgeführt werden (Eva et al., 2004 zitiert nach Harasym et al., 1996). Diese Problematik nehmen Eva und Kollegen zum Anlass eine neue Form der überfachlichen Kompetenz-Messung zu entwickeln: Multiple Mini-Interviews (Eva et al., 2004).

Das Vorbild für diese spezielle Form des Performance Assessment stammt ebenfalls aus der Medizin und nennt sich Objective Structured Clinical Examination (OSCE; Cleland et al., 2012). Damit ist eine Form des Abschlussexamens für die medizinische Ausbildung gemeint, in der die Absolvent*innen anhand konkret definierter Aufgaben praktisch beweisen müssen, dass sie das medizinische Handwerkszeug erworben haben. Beispielsweise müssen sie

eine Anamnese erheben, eine Ultraschalluntersuchung durchführen oder Patient*innen bestimmte Befunde vermitteln. Diese praktische Prüfung ergänzt häufig die schriftlichen Prüfungen.

Das MMI unterscheidet sich vom OSCE nun dadurch, dass nicht die Ergebnisse der medizinischen Ausbildung, sondern deren Voraussetzungen geprüft werden. Damit kann in den Problemsituationen kein medizinisches Wissen vorausgesetzt werden. Dafür werden persönliche Dispositionen wie Empathie, soziale Kompetenz und Argumentationsvermögen erfasst.

Im Unterschied zu einem umfassenden Interview mit mehreren Beurteilenden wird im MMI jede Situation von einer einzelnen Person bewertet. Damit sind die Urteile aller Situationen voneinander unabhängig und die Kandidat*innen können jede Situation mit neuer Zuversicht in Angriff nehmen. Da die einzelnen Beurteiler*innen nur eine einzelne Situation betrachten, haben sie gute Vergleichsmöglichkeiten hinsichtlich der Kandidat*innen und können auch entsprechende Expertise im Hinblick auf die Einschätzung des gezeigten Verhaltens entwickeln. Separate Beurteilungen kurzfristig angelegter Leistungssituationen sind somit diagnostisch ergiebiger als die globale Bewertung eines Interviews oder einer mündlichen Prüfung mit vielen unterschiedlichen Inhalten.

■ Empirische Befunde zu Akzeptanz und Brauchbarkeit

Die bestehende Literatur zeigt, dass das MMI sowohl von Seiten der Bewerbenden als auch von Seiten der Beobachtenden mit einer meist überdeutlichen Mehrheit als fair und akzeptabel angesehen wird (z.B. Brownell et al., 2007; Humphrey et al., 2008; Kelly et al., 2014b). Zudem empfinden mehr als 80 % der Teilnehmenden einer Untersuchung, dass das MMI ihnen geholfen hat, die eigenen Stärken zu präsentieren (Campagna-Vaillancourt et al., 2014). 90 % einer anderen Studie sind der Meinung, dass die Inhalte eines MMIs relevant für ihr Verständnis der medizinischen Praxis sind, wohingegen diese Meinung nur von 60 % der Proband*innen, die ein klassisches Interview absolviert haben, vertreten wird (Kelly et al., 2014a). Selbst abgelehnte Bewerbende beurteilen das Verfahren überwiegend positiv (Razack et al., 2009). Auch im statistischen Vergleich mit einem klassischen Interview wird das MMI als signifikant fairer, effektiver, aber auch stressintensiver beurteilt (Humphrey et al., 2008; Razack et al., 2009). Kritikpunkte des Verfahrens beziehen sich neben der empfundenen Stressintensität auf beiden Seiten und einer hohen Beanspruchung auf Seiten der Beobachtenden zudem auf die Möglichkeit, sich durch die MMI-Situationen schauspielern zu können (Razack et al., 2009). Auch die Trainierbarkeit mancher Situationen wird angemerkt; allerdings zeigen sich Effekte erst ab einem sehr hohen Vorbereitungsaufwand (Laurence et al., 2013).

■ Erkenntnisse zu qualitätsbestimmenden Merkmalen

Objektivität

Die besondere Stärke des MMI-Formates ist dessen hohe Durchführungsobjektivität. Die bei den einzelnen Stationen zu bearbeitenden Aufgabenstellungen werden schriftlich präsentiert und die Zeit für die Vorbereitung ist genau festgelegt. Bei vielen Stationen betreten die Proband*innen den Testraum und beginnen unmittelbar mit ihren Darlegungen. Wenn Rollenspielpartner*innen vorgesehen sind, verhalten sich diese in der Regel nach einem festen Drehbuch, oft mit standardisierten Äußerungen. Auch die Auswertungsobjektivität wird in

der Regel durch verschiedene Maßnahmen gesichert. Im Mittelpunkt steht ein standardisiertes Beobachtungs- oder Beurteilungsraster mit vorgegebenen Kategorien und Ausprägungen. Checklisten- oder Ratingskalen, d. h. gebundene Verfahren, sind die Mittel der Wahl, freie Notizen werden allenfalls ergänzend vorgenommen.

Durch die feste Zuordnung von einzelnen Beobachtenden zu bestimmten Interview-Stationen ist es möglich, Beobachter-Effekte zu minimieren und teilweise zu egalisieren. Da jede*r Bewerbende in jeder Situation auf dieselben Beobachtenden trifft, ist die Beurteilung unabhängig von der individuellen Beurteilungsweise (Tendenz zur Milde/Mitte/Härte). Jede*r Teilnehmende erfährt dieselbe Strenge in derselben Situation und erlebt somit denselben Schwierigkeitsgrad. Zudem bedeutet jede Interview-Station einen neuen Start für die Bewerbenden, unbeeinflusst von den vorangegangenen Leistungen, da diese von dem neuen Stationspersonal nicht gesehen wurden. Müssen aufgrund hoher Bewerbendenzahlen verschiedene Beurteilende an parallelen Stationen eingesetzt werden, so ergibt sich die unterschiedliche Strenge bzw. Milde durchaus als Problem. Roberts et al. (2014) berichten, dass bis zu 40 % der Varianz der Beurteilungen auf die Maßstäbe der Beurteilenden zurückzuführen sind. Da es aber offenbar stabile persönliche Standards gibt, kann man diese mathematisch im Endergebnis so berücksichtigen, dass die Kandidat*innen der strengen Prüfer*innen im Vergleich zu den anderen nicht benachteiligt werden (Roberts et al., 2010).

Reliabilität

Hinsichtlich der Reliabilität müssen verschiedene Ansätze unterschieden werden. Da in vielen Verfahren pro Station nur ein Gesamtwert ermittelt wird, können – sofern an der Station mehrere Beurteilende im Einsatz sind – deren Übereinstimmung berechnet werden. So berichten Sebok et al. (2014) an einer Testpersonengruppe von 444 Personen, dass die Korrelationen zwischen den gleichzeitig agierenden studentischen und akademischen Beurteiler*innen zwischen .41 und .69 lagen. Allerdings lag es auch an der zu bearbeitenden Aufgabe, wie gut und übereinstimmend das Verhalten zu beurteilen war. So finden sich beispielsweise für die Beurteilungsdimension “Kritisches Denken“ Übereinstimmungswerte von .41 bis .65, je nach Situation.

Die Globalwerte pro Station können mit den an den anderen Stationen erzielten Werten auf ihre Übereinstimmung verglichen werden. Hier erzielte beispielsweise Callwood (2015) trotz überwiegend übereinstimmender Beurteilungsdimensionen pro Station nur geringe Korrelationen zwischen den acht Stationen (ähnlich Cox et al., 2015). Im Unterschied dazu fanden Gafni et al. (2012) relativ hohe Konsistenzschätzungen über 14 Stationen von etwa .70, auf der Basis hoher Probandenzahlen (über 4000 verteilt auf mehrere Jahrgänge). Abdul Rahim und Yusoff (2016) berichten einen Reliabilitätskoeffizienten von .94 bei der Verwendung von sechs identischen Beurteilungsdimensionen über fünf Stationen. Dowell et al. (2012) fanden in mehreren Durchgängen Reliabilitätswerte von etwa .85 bei 4 bis 6 Stationen für die Gesamtwerte pro Station, aber deutlich niedrigere Werte für spezifische Merkmale wie kritisches Denken (~ .35), moralisches Urteil (~ .30) oder Teamwork (~ .15).

Werden an den Stationen mehrere Verhaltensdimensionen beurteilt, so kann deren Konsistenz pro Station bestimmt werden. Callwood (2015) erzielte hier sehr hohe Werte zwischen .91 und .97, unter Verwendung relativ allgemeiner persönlichkeitsbezogener Einschätzdimensionen. Ähnlich hohe Werte berichten Cox et al. (2015).

In einer Pilotstudie konnte Eva und Kollegen (2004) zeigen, dass die Erhöhung der Interview-Stationen mit jeweils einer/einem Beobachtenden einen größeren Einfluss auf die Reliabilität hatte als die Anhebung der Zahl von Beobachtenden in einem Interview (statt $r = .55$ mit 12 Beobachtenden $r > .81$ bei sechs oder mehr Stationen).

Bis einschließlich 2014 existierten 40 Studien, die die Reliabilität von MMIs prüften. Überwiegend zeigen sich akzeptable bis sehr gute Werte. Allerdings wird in gut der Hälfte der Studien ebenfalls die Varianz, die auf Bewerber-Unterschiede zurückzuführen ist, analysiert. Dabei zeigt sich, dass die aufgeklärte Varianz über die Studien hinweg zwischen 10 und 74 % variiert, wobei sie in den meisten Fällen weniger als 30 % der Gesamtvarianz ausmacht (Knorr & Hissbach, 2014). Neben der Erhöhung der Anzahl von Stationen wirken sich zudem folgende Faktoren positiv auf die Reliabilität aus: Ausschluss von sehr leichten Interview-Stationen sowie ein normativ verankertes Beurteilungssystem, d. h. eines mit möglichst wenig subjektivem Beurteilungsspielraum (Uijtdehaage, Doyle & Parker, 2011).

Validität

Die Beurteilung der Validität von MMIs gestaltet sich vergleichsweise schwierig, da das Verfahren noch sehr jung ist. Es zeigt sich häufig, dass die MMI-Ergebnisse mit traditionellen Eignungsindikatoren wie etwa der Durchschnittsnote des Hochschulzugangs nur gering korrelieren (z. B. Dowell et al., 2012; Simmenroth-Nayda et al., 2014). Dies ist aber durchaus beabsichtigt, da das neue Verfahren bislang nicht berücksichtigte Merkmale der Bewerbenden erfassen will.

Beim Einsatz verschiedener Verfahren zur Erfassung derselben Persönlichkeitsmerkmale oder Kompetenzen (z. B. eines Situational Judgment Test und eines MMI-Verfahrens), lassen sich mittelhohe Korrelationen finden (Roberts et al., 2014). Oliver et al. (2014) berichten Korrelationen um $r = .40$ für kommunikative Kompetenz und Problemlösen, jeweils erfasst durch MMI-Situationen und alternativ durch thematische Interviews.

Verschiedene Publikationen über das McMaster-MMI verdeutlichen dessen inkrementelle Validität im Rahmen des kombinierten Zulassungsverfahrens (Durchschnittsnote und MMI). Für andere MMIs gilt dies allerdings nicht. Dennoch lässt sich anhand der Arbeit von Knorr und Hissbach (2014) aufzeigen, dass MMIs vorwiegend keine kognitiven Fähigkeiten erfassen, weil die Zusammenhänge mit entsprechenden Messverfahren über verschiedene Studien hinweg bestenfalls gering ausgeprägt sind. Stehen intellektuelle Fähigkeiten im Hintergrund, besteht dagegen ein geringer bis moderater Zusammenhang. Daher eignen sich MMIs tendenziell dazu, ein ausschließlich auf intellektuellen Fähigkeiten basierendes Auswahlverfahren zu ergänzen. Kritisch zu beurteilen ist dabei die fehlende Konstruktvalidität der meisten MMIs (Patterson et al., 2016).

Die prädiktive Validität von MMIs ist insofern gegeben, als dass die Testwerte mit späteren Leistungsprüfungen im Studium signifikant korrelieren (Pau et al., 2013). Erwartungsgemäß fallen die Korrelationen besonders hoch zu praktischen Abschlussprüfungen aus, wobei (minderungskorrigierte) Werte bis zu $r = .50$ erreicht werden können (Husbands & Dowell, 2013). Dabei ist zu berücksichtigen, dass andere Maße in der Regel geringere Vorhersageleistungen erreichen. Dies spricht immerhin dafür, dass Kandidat*innen, die in den MMI-Prüfungen sehr schwach abschneiden, voraussichtlich auch im weiteren Studium eher schwach abschneiden, was einen Ausschluss entsprechender Personen mit extrem ungünstigen Werten

vom Studium erlauben würde. So können Heldenbrand et al. (2016) für das Pharmaziestudium zeigen, dass Probanden mit ungünstigen MMI-Werten zu Beginn des Studiums dreimal so oft schwache Studienergebnisse erzielten als Probanden mit besseren Werten. Patterson et al. (2016) berichten in ihrer Überblicksstudie insgesamt von wachsenden Belegen zwischen MMI-Daten und späteren, vor allem den praktischen Prüfungsergebnissen im Medizinstudium.

In einer Überblicksstudie über 11 Jahre MMI-Einsatz an der McMaster University Medical School weisen Rees et al. (2016) darauf hin, dass darauf zu achten sei, dass Bewerber*innen aus besonderen sozialen oder ethnischen Gruppen nicht benachteiligt würden. Hier gibt es in einigen Studien nämlich entsprechende Hinweise.

Kosteneffizienz

Im Anschluss an das erfolgreiche Pilotprojekt analysieren Eva und Kollegen (2004) die Kosten im Vergleich zu dem bis dato implementierten klassischen Interview. Dabei wird deutlich, dass im Besonderen mehr Bewerbende in kürzerer Zeit beurteilt werden können. Insgesamt werden so 50 % der bisherigen Kosten eingespart, da anstelle von vier Personenstunden pro Kandidat*in nur zwei benötigt werden. Sehr gering bemessen wurde dabei jedoch der Entwicklungsaufwand einer MMI-Situation (ca. drei Arbeitsstunden mit Gesamtkosten von 50 US$). Hissbach und Kollegen (2014) weisen darauf hin, dass bei einer theoretisch fundierten und gut evaluierten Situation die Kosten der Entwicklung erheblich höher sind: rund 2.000 US$ je Situation auf Basis von ungefähr 40 Arbeitsstunden. Zudem kann der enorme Unterschied dieser Angaben auch auf die Komplexität der einzelnen Stationen und bzw. deren Beurteilungssysteme zurückgeführt werden. So wird bei Eva und Kollegen (2004) beispielsweise nur die globale Einschätzung mithilfe einer siebenstufigen Likert-Skala vorgenommen. In diesem Fall wäre ein detailliertes Feedback, wie es beim Performance Assessment in Deutschland üblich ist undenkbar. Mehrdimensionale Verhaltensregistrierungen bzw. -beurteilungen sind zwar erheblich aufwändiger, machen eine Entscheidung aber auch deutlich transparenter. Zudem bieten sie die Möglichkeit eines differenzierten Feedbacks. Hierdurch könnte die Kompetenz- & Persönlichkeitsentwicklung der Teilnehmenden gefördert werden, nicht zuletzt weil, durch die Offenlegung von persönlichen Stärken und Schwächen, gezielte Entwicklungsmaßnahmen möglich sind.

Eigene Arbeiten

Seit dem Sommersemester 2015 arbeiten an der Universität Erfurt verschiedene Gruppen unter der Leitung des Zweitautors an der Entwicklung eines MMIs für die Lehramtsausbildung. In einem ersten Pilotprojekt haben zwei Studierende des Masterstudienganges Psychologie, darunter die Erstautorin, im Rahmen ihres Schwerpunktfaches acht verschiedene MMI-Situationen entwickelt und nach entsprechenden Pretests der einzelnen Situationen eine erste qualitative Erhebung des Konzeptes „MMI“ durchgeführt (Januar 2016). Im Rahmen ihrer Master-Thesis hat Frau Lüllemann die Arbeit aus ihrem Schwerpunkt fortgeführt und eine der Situationen mit Blick auf die Reliabilität evaluiert (Sommersemester 2016). Im Wintersemester 2016/17 entwickelte eine Gruppe von Psychologiestudierenden weitere MMI-Situationen und erprobte diese im Verlauf des Sommersemesters 2017. Im Vordergrund stand bei diesen Arbeiten erneut die Zuverlässigkeit der Einschätzungen.

■ Erste Entwürfe und Erprobungen 2015/16 (Lüllemann & Simeth, 2016)

In einem ersten Schritt wurden ähnlich einer Anforderungsanalyse diejenigen Kompetenzen ermittelt, die mithilfe des MMI-Verfahrens erfasst werden sollen. Die theoretischen Eckpfeiler hierfür bildeten die „Standards für die Lehrerbildung“ (KMK, 2004) und das „Hierarchische Strukturmodell von Handlungskompetenz“ (Frey, 2008). Auf Grundlage dessen wurden diejenigen Kompetenzen ausgewählt, die sowohl bei Frey als auch bei der KMK als Anforderungen an den Lehrberuf festgeschrieben sind. Weil das Ziel des MMI-Verfahrens die Messung von überfachlichen Kompetenzen ist, blieben die fachlichen Anforderungen unberücksichtigt. Schnittmenge des Vergleichs waren die nachstehenden Kompetenzen:

- Soziale Verantwortung,
- Gelassenheit und Geduld,
- Werte und Normen,
- Konfliktfähigkeit,
- Kommunikationsfähigkeit und strukturiertes Vorgehen,
- Problemlösefähigkeit unter belastenden Umständen,
- Reflexivität und Veränderungsbereitschaft (Kritikfähigkeit).

Davon ausgehend wurden acht Situationen entwickelt, die jeweils einen der Kompetenzbereiche erfassen sollen. Gleichzeitig wurde versucht darauf zu achten, dass ein Bezug zum Lehrberuf hergestellt wurde, um die Inhaltsvalidität und so auch die Akzeptanz des Verfahrens zu gewährleisten. Der Situationskatalog wurde daraufhin mit einer Gruppe von wissenschaftlichen Mitarbeitenden mit psychologischer oder Lehramtsausbildung beraten und optimiert. Mit den in Tabelle 1 knapp dargestellten MMI-Stationen wurde mit Ausnahme der Feedback-Situation jeweils ein Pretest auf Funktionalität der einzelnen Situationen durchgeführt.

Pretest

Ziel war es zu prüfen, ob die Situationen von Teilnehmenden verstanden werden und zudem das erwartete Verhalten ausgelöst wird. Hierfür erklärten sich zwei Lehramtsstudentinnen des Masterprogramms bereit. Ebenfalls sollten die situationszugehörigen Beurteilungsbögen auf Verständlichkeit und Nützlichkeit geprüft werden. Daher nahmen auch zwei Psychologie-Studentinnen mit entsprechendem Schwerpunktfach als zusätzliche Beobachterinnen am Pretest teil. Des Weiteren wurden die drei Schauspielrollen durch freiwillige Mitarbeitende und/oder Studierende der Universität Erfurt gestaltet. Nach jeder Situation wurden die Teilnehmerinnen im Rahmen eines teilweise standardisierten Interviews um ein Feedback gebeten. Die übrigen Beteiligten trafen sich sowohl zur Vorbereitung auf den Pretest als auch zur abschließenden Feedbackrunde nach dem Pretest.

Insgesamt konnte der Probedurchlauf als Erfolg gewertet werden. Die Situationen wurden von den Teilnehmerinnen als realistisch wahrgenommen und die Beurteilungsbögen spiegelten die tatsächlichen Beobachtungen wider. Allerdings wurde deutlich, dass die Bewertung des Verhaltens anhand einer Checkliste unbefriedigend ist, da qualitative Unterschiede bei der Bewertung nicht berücksichtigt werden können. Daher wurden die Beurteilungsbögen

überarbeitet, sodass eine möglichst objektive Bewertung der Performances auf einer vierstufigen Ratingskala erfolgen kann (Beispiel in Abbildung 2). Zudem zeigte sich Optimierungsbedarf in allen drei Schauspielsituationen. Hier konnte bei den Schauspielenden in der jeweils zweiten Begegnung eine zwar reduzierte, aber weiterhin präsente Unsicherheit festgestellt werden. Um dem künftig vorzubeugen und die Situationen gerechter zu gestalten, wurden entsprechend detaillierte Schauspielleitfäden entwickelt. Darüber hinaus waren die Teilnehmer*innen merklich irritiert und verunsichert, dass sie in der Dilemma-Situation keinen Interaktionspartner hatten. Demgemäß war die Situation in beiden Fällen schnell beendet und induzierte nicht das gewünschte Verhalten. Nach einer Anpassung des Situationsrahmens wurde dieses Mini-Interview erneut von zwei anderen Studierenden der Universität Erfurt absolviert – beide Male wie bezweckt.

Tabelle 1
Kurzbeschreibung der finalen MMI-Situationen

Situation	Kompetenz	Typ	Kurzbeschreibung
Klassenfahrt	Problemlösen unter belastenden Umständen	Rollenspiel	Die/der Bewerbende ist eine/r von zwei Betreuenden einer Klassenfahrt, die heute zu Ende geht. Vermeintliches Ziel ist es die Klassenfahrt mit der Kollegin zu evaluieren. Tatsächlich findet die/der Bewerbende eine aufgelöste Kollegin vor, weil zwei Kinder unauffindbar sind.
Bilder-Chaos	Kommunikationsfähigkeit	Monolog	Ein aufwendiges geometrisches Bild soll beschrieben werden, sodass eine andere Person, die das Bild nicht kennt, dieses nachzeichnen könnte.
Projektwoche	Konfliktfähigkeit	Rollenspiel	Eine uneinsichtige Lehrer-Kollegin hat sich für das bevorzugte und bereits vorbereitete Thema zur Projektwoche angemeldet, obwohl sie von dem Vorhaben der/des Bewerbenden wusste.
Motive	Soziale Verantwortung	Monolog	Bewerbende/r soll Herausforderungen im Lehrberuf, die eigene Motivation Lehrende/r zu werden und ihre/seine Eignung dafür beschreiben.
Hausmeister	Gelassenheit und Geduld	Rollenspiel	Aufgrund unglücklicher Umstände wird dringend ein neuer CD-Spieler benötigt. Die letzte Möglichkeit ist der gemächliche Hausmeister. Leider bleiben nur noch 5 Minuten bis die Unterrichtsstunde anfängt.
Dilemma	Werte und Normen	Monolog	Befreundete Eltern überlegen ihr Kind wegen befürchteter Hänseleien im Sportunterricht krank zu melden. Sie sollen als Lehrperson Rat geben, indem Sie Handlungsmöglichkeiten und deren Konsequenzen diskutieren.
Frustrierter Arzt	Kritikfähigkeit	Monolog	Aus einem kritischen Forenbeitrag sollen die einzelnen, polemisch formulierten, Kritikpunkte aufgegriffen und jeweils mit einer sachlichen Stellungnahme versehen werden.
Feedback*	Reflexivität und Veränderungsbereitschaft	Monolog	Im Anschluss an alle Situationen und nach einer kurzen Pause (Beobachterkonferenz) sollen die Bewerbenden ihre Performance im vorangegangenen MMI reflektieren.

*außerhalb des Rotationsverfahrens

Verhalten - Positiv	*gezeigt*	*nicht gezeigt*	*nicht beurteilbar*
Verwendet zur räumlichen Orientierung Richtungsangaben oben, unten, links, rechts, Winkelangaben			

Abbildung 1: Beispiel für einen Beurteilungsanker per Checkliste.

Verwendet zur räumlichen Orientierung Richtungsangaben	*0-1 von 4*	*2 von 4*	*3 von 4*	*4 von 4*
- oben, unten, links, rechts - Winkelangaben - Längenangaben - Werkzeuge	①	②	③	④

Abbildung 2: Beispiel für einen Beurteilungsanker per Ratingskala

MMI-Erprobung

Neben der Funktionalität der einzelnen Situationen war das übergeordnete Ziel die Erprobung einer regelrechten Durchführung eines MMIs. Hierfür benötigt man idealerweise zwei Beobachtende und eine*n Teilnehmer*in je Interview-Station, bei Rollenspielen zudem eine*n entsprechende*n Schauspieler*in und eine weitere Person, die für die korrekte Durchführung des MMIs Sorge trägt.

Für die hiesige Erprobung zeichnete sich früh ab, dass zum geplanten Erhebungszeitraum nicht genügend Beobachtende verfügbar waren, um die Interview-Stationen doppelt zu besetzen. Zudem musste berücksichtigt werden, dass die verfügbaren Räumlichkeiten teilweise sehr klein waren. Insgesamt beteiligten sich daher sieben wissenschaftliche Mitarbeitende des Zweitautors als Beobachtende, weitere sieben Lehramtsstudierende des Masterprogramms als Teilnehmende, die drei Schauspielenden aus dem Pretest für jeweils dieselben Rollen sowie die zwei Organisatorinnen des Verfahrens an der MMI-Erprobung. Um das Prozedere besser nachvollziehen zu können, werden im Folgenden die Organisation sowie die Durchführung der Erprobung geschildet.

Nach der Akquise aller Beteiligten wurden in einem ersten Schritt die notwendigen Materialien an die jeweils Betroffenen versendet: Die Lehramtsstudierenden erhielten vorab ein Anschreiben, welches sie auf das bevorstehende MMI vorbereiten sollte. Hier wurden Aufgabentypen, der spezielle Charakter der MMI-Situationen sowie der wiederkehrende Ablauf erläutert. Auch die Beobachtenden erhielten ein allgemeines Informationsblatt, um jede*n zu einer/einem Spezialist*in dieses MMIs zu machen. Darüber hinaus war jede*r Beobachtende als Expert*in für eine Situation vorgesehen. Daher wurde die entsprechende Aufgabenstellung mit situationsspezifischen Beobachter-Instruktionen und Beurteilungsbögen ebenfalls im Vorfeld an die entsprechenden Adressaten verschickt. Zudem wurden sie zu einer kurzen Beobachterschulung geladen, in der insbesondere letzte offene Fragen beantwortet werden sollten. Auch die Schauspielenden erhielten frühzeitig ein entsprechendes Informationsblatt sowie ihren Schauspiel-Leitfaden und konnten in einer kurzen Auftaktveranstaltung letzte Unsicherheiten beseitigen. Um ebenfalls valide Aussagen zur Erprobung machen zu können, rotierte eine der beiden Organisatorinnen durch die einzelnen MMI-Stationen in

entgegengesetzter Richtung. So war sichergestellt, dass jede Situation, von Seiten der Organisatorinnen, zumindest einmal beobachtet werden konnte und gleichzeitig unterschiedliche Persönlichkeiten Teil dieser Beobachtungen waren.

Abbildung 3 soll den Ablauf der eigentlichen Erprobung visualisieren und die Verlängerung des Verfahrens durch die Feedback-Situation im Anschluss an das Rotationsverfahren verdeutlichen. Zur Beurteilung dieser Interview-Station war es notwendig, dass die Beobachtenden sich über die unterschiedlichen Teilnehmenden und deren Performances austauschten. Eine zusätzliche Schwierigkeit war dabei der Wechsel der Bezugsgruppen: Anstelle eines Situationsexperten musste in kürzester Zeit jede*r Beobachtende Expert*in für eine*n Teilnehmende*n werden. Dies war insbesondere in Bezug auf den beratenden Charakter des Teaching Talent Centers interessant und sollte einen ersten Eindruck zur Durchführbarkeit von Feedback an die Bewerbenden vermitteln. Neben dieser Herausforderung zeigte sich zudem, dass der Appell zur Selbstreflexion teilweise maximal drei Sätze umfasste und erst in der Interaktion mit der/dem Beobachtenden ergiebiger wurde. Im Anschluss an diese letzte MMI-Situation wurden die Teilnehmenden jeweils von ihrer/ihrem persönlichen Beobachtenden mithilfe eines teilweise standardisierten Interviews um ein Feedback zu dem Verfahren aus ihrer Sicht gebeten.

Abschließend wurde sowohl mit den Schauspielenden als auch mit den Beobachtenden über das Verfahren beraten. Die einhellige Meinung von allen Beteiligten war, dass die einzelnen Situationen realistisch und repräsentativ sind. Zudem kann festgehalten werden, dass das Verhalten der Teilnehmenden Qualitätsunterschiede aufzeigt und diese Varianz durch die Beurteilungsanker abbildbar ist. Teilweise müssen diese jedoch weiter konkretisiert und ausdifferenziert werden (z.B. Situation Projektwoche). Neben den situationsspezifischen Beobachtungen sollten auch situationsübergreifende Kompetenzen (z.B. soziales Interaktionsverhalten) beurteilt werden. Diese konnten nach Aussage der Beobachtenden anhand der Beobachtungsbögen nicht reliabel gemessen werden, da mehrere beobachtbare Indikatoren gleichzeitig als Einstufungskriterien herangezogen wurden. Die Rollenspiele werden dagegen als besonders positiv wahrgenommen, gehen aber mit der Problematik einher, dass teilweise die Situationen durch Erfindungen aufgelöst („ich habe die Kinder auf der Toilette gefunden“) oder als Gelegenheit zum tatsächlichen Schauspielern genutzt wurden. Um diese Gefahr zu eliminieren, soll künftig die Bewerbenden-Instruktion angepasst sowie das Wort „Rollenspiel“ durch „Interaktionssituation“ ersetzt werden.

Insgesamt wird deutlich, dass die Entwicklung und Durchführung eines MMIs mit hohem Aufwand verbunden ist, der möglicherweise durch den Ertrag des Verfahrens aufgewogen wird. Festzuhalten ist, dass die Resonanz aller Beteiligten positiv und die Akzeptanz dieses Verfahrens gewährleistet ist. Die hochgradig standardisierte Struktur und die damit sichergestellte Objektivität ist eine herausragende Stärke des MMIs. Da sich das Erfurter MMI-Verfahren zu diesem Zeitpunkt noch am Anfang der Erprobungsphase befindet, sind allerdings valide Aussagen zu Reliabilität und Validität noch nicht möglich. Um hier Abhilfe zu schaffen, hat sich Janet Lüllemann in ihrer Masterarbeit (Lüllemann, 2016) mit einer vertieften Analyse der Reliabilität beschäftigt.

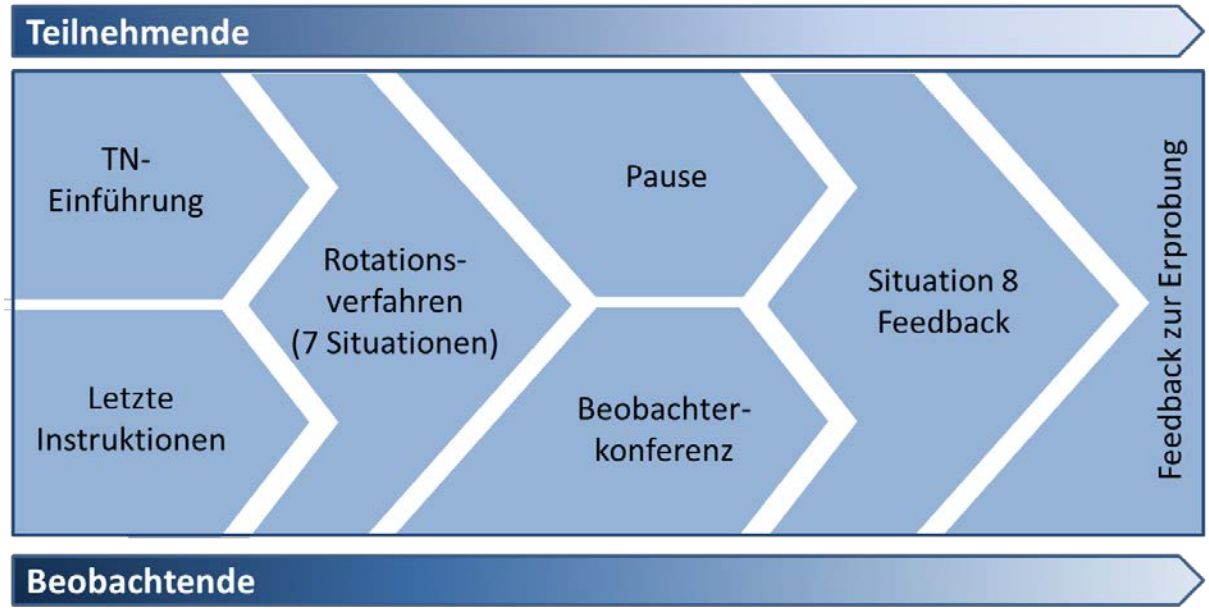

Abbildung 3: Ablauf der MMI-Erprobung.

■ Reliabilitätsprüfung des Bereichs „Instructional Clarity“ (Lüllemann, 2016)

Wahl des Inhaltsbereichs

Bei der Durchsicht der Aufgaben, die im Rahmen der Zulassungsverfahren für das Medizinstudium Verwendung fanden, fiel eine Aufgabe auf, die eine eindeutig pädagogische Komponente aufwies: Eine Testperson sollte eine abstrakte Grafik einer zweiten Person, die diese Grafik nicht sehen konnte, so beschreiben, dass die zweite Person in der Lage sei, diese Grafik möglichst korrekt zeichnerisch wiederzugeben. Dieser Aufgabentyp wird offenbar relativ häufig verwendet und soll im Bereich der medizinischen Eignungsdiagnostik die Fähigkeit zur Teamarbeit erfassen. Beispiele für die Gestaltung dieser Aufgabe finden sich im Internet (z. B. hier: http://bambai.xyz/info-GY70BrKGOdU.html).

Im Hinblick auf die Kompetenzrahmen für die Lehrerbildung fällt diese Aufgabe allerdings in den Bereich „Instructional Clarity“. Die Qualität des Unterrichts kann unter anderem daran gemessen werden, dass der im Unterricht behandelte Stoff vorausschauend vorgestellt wird, dass die Stoffpräsentation strukturiert erfolgt, dass Wechsel im Thema angekündigt werden und dass der Stoff abschließend zusammengefasst wird. Diese Vorgehensweisen fallen unter das Konzept „Instructional Clarity“ (Titsworth et al., 2015). Das übergeordnete Konzept der „Teacher Clarity“ wird bereits seit den siebziger Jahren in der pädagogischen Forschung behandelt. Es konnte gezeigt werden, dass dieses Konstrukt sowohl mit niedrig als auch mit hoch inferenten Skalen erfasst werden kann und dass die entsprechenden Indikatoren substantiell miteinander korrelieren (Hines, Cruickshank & Kennedy, 1985). Die Klarheit und Verständlichkeit der Präsentation hat Auswirkungen auf die Lernprozesse und Lernerfolge der Schüler*innen, aber ebenso auf ihre Lernmotivation und ihre Zufriedenheit mit dem Unterricht (Ribera et al., 2012).

Helmke (2009, S. 191ff.) setzt sich differenziert mit dem Konzept der „Klarheit und Strukturiertheit“ auseinander und unterscheidet hinsichtlich der Klarheit die akustische, die sprachliche, die inhaltliche und die fachliche Klarheit. Die *akustische Klarheit* bezieht sich auf die Verständlichkeit des Gesagten in Abhängigkeit von der Lautstärke, der Tonhöhe, der Sprechgeschwindigkeit, der Pausen usw., die der Sprecher bzw. die Sprecherin benutzt. Die *sprachliche Klarheit* bezieht sich auf die Korrektheit der Grammatik, die Vollständigkeit der Sätze, Verzögerungen, Phrasen und andere Formen des Sprechens, die die Sinnerfassung fördern oder beeinträchtigen. *Inhaltliche Klarheit* meint vor allem die Präzision der verwendeten Begriffe und die Kohärenz der Aussagen und Argumente. Hier sind oft auch die Strukturierungshilfen angesprochen, die mit Verweis auf Titsworth et al. oben bereits referiert wurden. Mit *fachlicher Klarheit* ist vor allem die Korrektheit der Ausführungen gemeint; dies kann ergänzt werden mit Blick auf die Verwendung von angemessenen Fachbegriffen im passenden Kontext.

Untersuchtes Material

Die Klarheit der unterrichtlichen Ausführungen kann sicherlich auf der Basis eines Lehrvortrags beurteilt werden; allerdings wird vielfach argumentiert, dass die Verständlichkeit von Erklärungen einen Dialog zwischen Lehrenden und Lernenden erfordert, sodass die Lehrperson auf Verständnisprobleme interaktiv eingehen kann (z. B. Simonds, 1997). Im Rahmen eines eignungsdiagnostischen Verfahrens erscheint es jedoch im Hinblick auf die Objektivität der Durchführung problematisch, eine oder mehrere Personen als Adressaten der Instruktion vorzusehen, die bei jeder Durchführung bestimmte Verständnisprobleme schauspielern müssten. Deshalb wurde nach Aufgabenformaten gesucht, die es erlauben sollten, ohne eine*n konkrete*n Lernpartner*in Lern- oder Handlungsanweisungen zu formulieren.

Zwei Aufgaben wurden konstruiert, die gleichzeitig eingesetzt werden sollten, um auch erstmalig die Frage der Konstruktstabilität überprüfen zu können.

Die *erste Aufgabe* bestand darin, eine Zeichnung mit geometrischen Symbolen in Form einer Tonbandaufzeichnung so zu beschreiben, dass diese das Nachzeichnen allein aufgrund der verbalen Anleitung ermöglichen sollte. Damit hatten die Kandidat*innen nur ein Audioaufzeichnungsgerät, aber keine konkrete Person vor sich. Diese Aufgabe wurde damit begründet, dass geübt werden sollte, auditive Instruktionsmaterialien zu erstellen. Diese Aufgabe wird im folgenden *Bildbeschreibung* genannt. Abbildung 4 zeigt ein Symbolbild (aus Testschutzgründen nicht die Originalaufgabe).

Die *zweite Aufgabe* bestand darin, anhand des Stadtplans einer fiktiven Stadt einer angeblich unerfahrenen Kollegin den Weg von der Schule zu einem Museum zu beschreiben, wobei die Beschreibung von einem Anrufbeantworter aufgezeichnet wurde, da die Kollegin zum Zeitpunkt des Anrufs nicht persönlich erreichbar war. Diese Aufgabe wird im folgenden *Wegbeschreibung* genannt. Abbildung 5 zeigt eine der präsentierten Vorlagen mit den Straßennamen. Eine weitere Abbildung enthielt alternativ die Bezeichnungen der Gebäude. Beide Abbildungen durften verwendet werden.

Beide Aufgaben erforderten demnach die sequenzielle Beschreibung von miteinander zusammenhängenden Handlungen gegenüber einer unbekannten Person, wobei mögliche Verständnisschwierigkeiten vorweggenommen und durch präzise und anschauliche Beschreibungen ausgeräumt werden sollten.

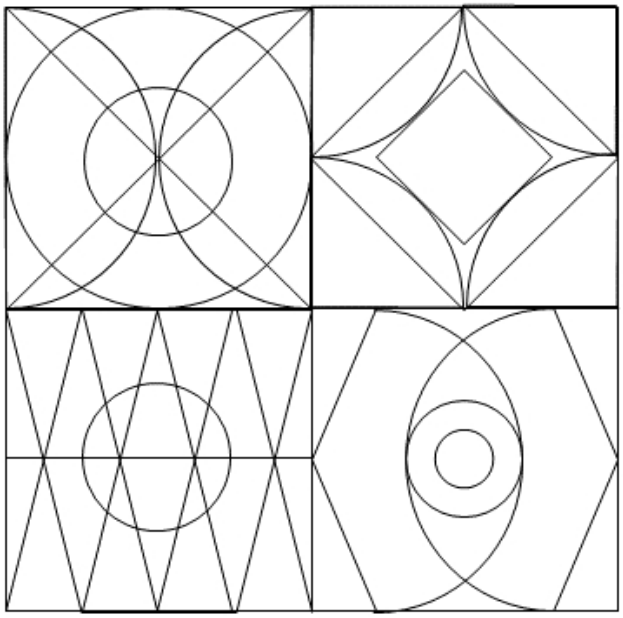

Abbildung 4: Symbolbild zur Aufgabe „Bildbeschreibung“.

Abbildung 5: Originaltestmaterial zur „Wegbeschreibung“; der Weg von der Schule (im Südosten) zum Museum (im Nordwesten) sollte verbal beschrieben werden. Eine zweite Karte enthielt die Bezeichnungen der Gebäude.

Untersuchungsfragestellung

Lüllemann (2016) verfolgte in ihrer Studie unter anderem folgende Untersuchungsfragestellungen:

- FS1: Aufgrund der höheren Anschaulichkeit werden bei der Wegbeschreibung mehr Punkte erzielt als bei der Bildbeschreibung.
- FS2: Die bei der niedrig inferenten Einschätzung erzielten Punkte korrelieren mit den bei der hoch inferenten Einschätzung erzielten Punkten.
- FS3: Im Sinne der Konstruktvalidität korrelieren die bei der Wegbeschreibung erzielten Punkte mit den bei der Bildbeschreibung erzielten Punkten.
- FS4: Personen mit pädagogischer Ausbildung erzielen bei den Aufgaben mehr Punkte als Personen ohne pädagogische Ausbildung.
- FS5: Personen mit pädagogischer Ausbildung differenzieren bei ihrer Beurteilung stärker als Personen ohne pädagogische Ausbildung zwischen Personen mit und ohne pädagogischer Ausbildung.
- FS6: Die niedrig inferenten Einschätzungen erfolgen zuverlässiger als die hoch inferenten Einschätzungen.

Methode

Für die Beantwortung der Fragestellungen war es nötig, zu den beiden beschriebenen Aufgabenstellungen verschiedene Beurteilungsinstrumente zu erstellen. Bei dem Instrument mit den *niedrig inferenten* Einschätzungen wurden Merkmale des gezeigten Verhaltens relativ verhaltensnah im Sinne einer Checkliste registriert. So waren unter der Rubrik „Akustische Verständlichkeit“ die drei Merkmale „spricht in einer angemessenen Lautstärke“, „spricht in einem angemessenen Sprechtempo“ und „setzt angemessene Pausen“ als gegeben versus nicht gegeben zu markieren. Insgesamt bestand dieses Instrument aus 14 kategorialen Merkmalen, die den vier Formen der Verständlichkeit im Sinne von Helmke zugeordnet waren.

Das Instrument mit den *hoch inferenten* Einschätzungen bestand nur aus vier Ratingskalen, d. h. jeweils einer zu den vier Formen der Verständlichkeit. Auf einer vierstufigen Skala von „sehr schwach“ bis „sehr stark“ sollte das Ausmaß der Verständlichkeit beurteilt werden.

Ferner erforderte das Design, dass die Aufgabe von Personen mit bzw. ohne pädagogische Vorbildung durchgeführt wurde. Deshalb wurden im Kreise der Mitarbeiterinnen und Mitarbeiter des Lehrstuhls des Zweitautors sowie der Erfurt School of Education bzw. im Bekanntenkreis von Frau Lüllemann Personen mit entsprechenden Merkmalen gebeten, die Aufgaben durchzuführen und ihre Lösungen auf Tonband aufzuzeichnen. Aus den dafür gewonnenen Personen wurden drei Probanden mit professioneller pädagogischer Vorerfahrung (das heißt in der Regel mit Lehramtsausbildung und Lehrtätigkeit) und drei Probanden ohne professionelle Vorerfahrung ausgewählt. Zur Abgrenzung von den Beurteiler*innen werden diese Personen im Weiteren als „pädagogische Performer“ bzw. „nicht-pädagogische Performer“ bezeichnet.

Des Weiteren war erforderlich, die Verhaltensleistungen der Performer von pädagogisch erfahrenen bzw. pädagogisch unerfahrenen Beurteiler*innen bewerten zu lassen. Insgesamt konnten für diese Aufgabe neun Personen mit langjähriger Lehrerfahrung und 18 Studierende aus dem Bachelorstudiengang Psychologie (ohne Lehrerfahrung) gewonnen werden. Diese Personen werden im Weiteren als „Rater“ bezeichnet.

Um die Belastung der Rater in Grenzen zu halten, wurden jedem Rater nur die Leistungen von zwei Performern vorgegeben, und zwar jeweils von einem pädagogischen und einem nicht-pädagogischen Performer. Jeder Rater beurteilte die Verhaltensleistungen der beiden bearbeiteten Aufgaben und dies sowohl mit dem niedrig inferenten als auch mit dem hoch inferenten Instrument.

Insgesamt lagen am Ende für jede Verhaltensleistung Beurteilungen von neun Ratern vor. Für die Berechnung der mittleren Einschätzungen und für die Korrelationen zwischen den Einschätzungen wurde ein Datensatz aus sechs Performances mal neun Beurteilungen, d. h. 54 Fällen gebildet. Diese Fälle sind nicht vollständig unabhängig voneinander, was bei der Interpretation der Werte zu berücksichtigen ist.

Die Erhebungen wurden jeweils individuell und in ruhiger Umgebung durchgeführt. Die Testanweisung für die Performer erfolgte in standardisierter Form. Die Rater hatten die Möglichkeit, die auditiven Aufzeichnungen mehrfach anzuhören. Im Einzelfalle traten bei den Aufzeichnungen technische Probleme auf, die die Verständlichkeit des Aufgezeichneten beeinträchtigten. Dieses Material wurde von der weiteren Verwendung ausgeschlossen.

Ergebnisse

Zur Überprüfung der Fragestellung FS1 (vgl. S. 22) wurde bei den niedrig inferenten Einschätzungen der durchschnittliche Anteil der erreichten Checklistenpunkte berechnet. Bei der Bildbeschreibung ergab sich ein Mittelwert von 0.53 (SD=0.22), bei der Wegbeschreibung von 0.64 (SD=0.18). Die Differenz ist signifikant (t=-4.13; $p < .001$). Bei den hoch inferenten Einschätzungen wurde der durchschnittliche Anteil der Ratingpunkte berechnet. Hier wurden bei der Bildbeschreibung 2.72 Punkte (SD=0.67) und bei der Wegbeschreibung 3.11 Punkte (SD=0.55) erzielt. Auch diese Differenz ist signifikant (t=-4.23; $p < .001$). Wie vermutet, ist die Wegbeschreibung mit Hilfe der Landkarten leichter als die Bildbeschreibung der geometrischen Muster.

Zur Überprüfung der Fragestellung FS2 wurden die Ergebnisse der hoch bzw. niedrig inferenten Ratings miteinander korreliert. Bei der Bildbeschreibung ergab sich ein $r = .83$ ($p < .001$), bei der Wegbeschreibung ein $r = .70$ ($p < .001$). Wie bereits von Hines et al. (1985) festgestellt wurde, erbringen die verschiedenen Skalierungen somit ähnliche Ergebnisse.

Zur Überprüfung der Fragestellung FS3 wurden die Einschätzungen zwischen den beiden Aufgaben korreliert. Auf der Basis der 54 Fälle ergab sich bei den Summenwerten der niedrig inferenten Einschätzungen ein $r = .54$ ($p < .001$) und bei den Summenwerten der hoch inferenten Einschätzungen ein $r = .40$ ($p < .005$). Die Hypothese ist damit bestätigt. Allerdings scheinen die beiden Aufgaben nur teilweise dasselbe Konstrukt zu erfassen.

Zur Überprüfung der Fragestellung FS4 wurden die 27 Einschätzungen für die pädagogischen Performer mit den 27 Einschätzungen für die nicht-pädagogischen Performer verglichen. Die

vier Summenwerte wurden jeweils per *t*-Test verglichen. Dabei zeigte sich, dass die pädagogisch geschulten Performer bei der geometrischen Bildbeschreibung signifikant besser abschnitten als die ungeschulten Performer (siehe Tabelle 2). Bei der Wegbeschreibung via Stadtplan zeigte sich dieser Unterschied jedoch nicht. Die anspruchsvolle schulnahe Aufgabe der Bildbeschreibung scheint also pädagogische Kompetenzen valide zu erfassen.

Tabelle 2
Vergleich der Mittelwerte der pädagogischen bzw. nicht-pädagogischen Performer

	nicht-pädagogisch		pädagogisch			
	M	SD	M	SD	*t*-Wert	*p*
Bild: niedrig inferent	0,43	0,17	0,64	0,20	-4,11	.00
Bild: hoch inferent	2,44	0,60	3,00	0,62	-3,39	.00
Weg: niedrig inferent	0,63	0,18	0,65	0,19	-0,52	.60
Weg: hoch inferent	3,06	0,63	3,15	0,47	-0,55	.58

Für die Überprüfung der Fragestellung FS5 (vgl. S. 22) wurde eine Varianzanalyse für jeden Summenwert durchgeführt. Neben dem pädagogischen Status der Performer (als within-subject-Variable für die Rater) wurde auch der pädagogische Status der Rater (als between-subject-Variable) berücksichtigt. In keinem Fall zeigte sich der erwartete signifikante Interaktionseffekt. Pädagogische Rater sind demnach nicht notwendig, um die besseren Leistungen pädagogisch geschulter Performer angemessen zu erkennen.

Für die Überprüfung der Fragestellung FS6 wurden für jedes Beurteilungsitem (d. h. Checklistenitem oder Ratingskala) zur Abschätzung der Reliabilität sogenannte Intraclass-Korrelationskoeffizienten (ICCs) berechnet. Diese Koeffizienten beruhen auf Varianzanalysen, bei denen die Performer als Fälle und die Rater als Variablen (d. h. Messwiederholungen) betrachtet werden. Systematische Unterschiede zwischen den Performern sind erwünscht, systematische Unterschiede zwischen Ratern (im Sinne von Milde- und Strengetendenzen) werden hingenommen. Problematisch sind Interaktionseffekte, die für eine unterschiedliche Rangreihe der Performer bei den einzelnen Ratern und damit für eine problematische Reliabilität der Skala sprechen. Somit können aus den mittleren Quadratsummen für die unterschiedlichen Varianzanteile Koeffizienten gebildet werden, die von ihrer Höhe her wie klassische Konsistenzkoeffizienten, z. B. Cronbachs alpha, interpretiert werden können (Shrout & Fleiss, 1979).

Die Koeffizienten wurden nach dem von Stemmler und Margraf-Stiksrud (2015) empfohlenen Vorgehen berechnet und so bestimmt, dass der Tatsache der wechselnden Rater Rechnung getragen wurde. Demnach wurde für einen durchschnittlichen Einzel-Rater der Wert ICC(1,1) und für die Gruppe von maximal neun Ratern der Gesamtwert ICC(1,9) berechnet. Dabei gibt der Wert ICC(1,1) Auskunft darüber, wie gut die Reliabilität des Items bei einem einzelnen Rater ist, während der Wert ICC(1,9) besagt, wie zuverlässig die Aussage bei einer Durchschnittsbildung über jeweils neun Rater wäre. Im Einzelfall musste die Zahl von 9 reduziert werden, nämlich dann, wenn ein Rater alle Performer mit demselben Wert beurteilt hatte.

Tabelle 3 zeigt die Werte für alle verwendeten Items sowie für die vier gebildeten Summenwerte.

Tabelle 3: *ICC-Werte (6 Performer, maximal 9 Rater) zur Abschätzung der Reliabilität der Items und Summenwerte*

	Bild	**ICC(1,1)**	**ICC(1,9)**	**Weg**	**ICC(1,1)**	**ICC(1,9)**
	niedrig inferent			**niedrig inferent**		
AV	Lautstärke	-	-	Lautstärke	-	-
	Sprechtempo	.18	.64	Sprechtempo	.18	.63
	Pausen	.15	.59	Pausen	.02	.14
SV	eindeutige Begriffe	.27	.77	eindeutige Begriffe	0	0
	Schwierigkeitsgrad	.16	.64	Schwierigkeitsgrad	.10	.49
	strukturiert	.20	.60	strukturiert	.09	.44
	ohne Korrektur	.10	.46	ohne Korrektur	0	0
IK	Ziel nennen	.40	.86	Ziel nennen	.38	.83
	Schritt-für-Schritt	.12	.50	Schritt-für-Schritt	.27	.52
	Hinweis auf Schwierigkeiten	.85	.97	Hinweis auf Schwierigkeiten	.37	.82
	ausführliche Beschreibung	.51	.89	ausführliche Beschreibung	.31	.78
FK	geometrische Figuren	-	-	Straßennamen	-	-
	Winkelangaben	.27	.75	Sehenswürdigkeiten	-	-
	Längenangaben	.24	.74	Eigenschaften der Gebäude	.49	.90
	hoch inferent			**hoch inferent**		
AV	akustische Verständlichkeit	.15	.61	akustische Verständlichkeit	.01	.11
SV	sprachliche Verständlichkeit	.36	.84	sprachliche Verständlichkeit	.02	.19
IK	inhaltliche Klarheit	.47	.98	inhaltliche Klarheit	0	0
FK	Nutzt Information aus Material	.29	.79	Nutzt Information aus Material	0	0
	Summenwerte			**Summenwerte**		
	niedrig inferent	.52	.91	niedrig inferent	.14	.59
	hoch inferent	.49	.89	hoch inferent	0	0

Legende: AV = Akustische Verständlichkeit; SV = Sprachliche Verständlichkeit; IK = Inhaltliche Klarheit; FK = Fachliche Klarheit; - = konnte wegen zu geringer Varianz zwischen den Performern nicht berechnet werden.

Die ICC-Werte zeigen, dass es bei den gestellten Aufgaben und den verwendeten Auswertungsschemata nicht möglich ist, durch eine einzelne Beurteilung ein zuverlässiges Resultat zu erzielen. Nur in einem Fall (Bildbeschreibung: Hinweis auf Schwierigkeiten) liegt der berechnete Wert über der Schranke von 0.8, die als notwendig für eine präzise Beurteilung angesehen wird. Interessanterweise erreichen auch die Summenwerte (am Ende der Tabelle) nicht die erforderliche Höhe. Ausgehend von einer größeren Beurteilergruppe, bessert sich das Bild in einigen Fällen: Von den 14 niedrig inferenten Kriterien der Bildbeschreibung überschreiten immerhin drei die genannte Schranke, bei den vier hoch inferenten Kriterien sind

es nahezu drei. Bei der Wegbeschreibung sind es ebenfalls drei der 14 niedrig inferenten Kriterien, die relativ zuverlässig durch eine größere Beurteilergruppe einzuschätzen sind; überraschenderweise gilt dies aber nicht für die hoch inferenten Einschätzungen.

Hinsichtlich der praktischen Brauchbarkeit erweist sich somit nur die Bildbeschreibung als ausreichend reliabel, und auch nur, wenn die einzelnen Leistungen von mehreren Ratern bewertet werden. Abstraktere (hoch inferente) Einschätzungen führen unerwarteter Weise zu stabileren Beurteilungen als konkretere (niedrig inferente) Einschätzungen.

Diskussion

Die von Lüllemann (2016) durchgeführte Studie erbrachte interessante Ergebnisse zur Reliabilität und Validität von MMIs. Da die Leistungen von beiden Aufgaben miteinander korrelierten, spricht dies für die Konstruktvalidität der Aufgaben; da pädagogische Fachleute bessere Leistungen erzielten als pädagogische Laien, spricht dies für die Kriteriumsvalidität. Beide Befunde zeigen in einer ersten Annäherung, dass mit den MMIs pädagogisch relevante Konstrukte erfassbar sind. Die Befunde zur Reliabilität sind hingegen ernüchternd. Die in der Praxis vorherrschende Vorgehensweise, dass die an einer MMI-Station gezeigte Leistung nur von einer Person und nur summarisch beurteilt wird, ist wohl nicht geeignet, eine zuverlässige Leistungseinschätzung hervorzubringen. Empfehlenswert ist auf jeden Fall der Einsatz mehrerer Beurteiler*innen, was auch die sehr umfassende Studie zur Unterrichtsbeurteilung der Bill und Melinda Gates-Stiftung (Ho & Kane, 2013) nahelegt. Es erweist sich als schwierig, verschiedene Verhaltensfacetten präzise kleinteilig zu registrieren, was aber für eine differenzierte Rückmeldung des gezeigten Verhaltens erforderlich wäre. Offenbar ist eine Globaleinschätzung des pro Station gezeigten Verhaltens einfacher und zuverlässiger als komplexere Alternativen. Eine aktuelle deutsche Studie (Praetorius et al., 2014) zeigt, dass es stark von der beobachteten Verhaltensdimension abhängt, wie viele Beurteilungen notwendig sind, um zu einer zuverlässigen Einschätzung zu kommen. Bei der Evaluation von Unterricht genügt beispielsweise eine einzelne Beobachtung (ein Unterrichtsausschnitt), um Klassenmanagement und kognitive Unterstützung beurteilen zu können. Soll hingegen einigermaßen zweifelsfrei einschätzen werden, ob Lehrkräfte ihre Schüler*innen kognitiv aktivieren können, sind bis zu neun Beobachtungen erforderlich.

Einschränkend muss gesagt werden, dass in der Studie von Lüllemann (2016) nur akustische Aufzeichnungen und eine begrenzte Zahl von Performances vorlagen. Außerdem erwies sich die Aufgabe der Wegbeschreibung als relativ einfach, sodass sich die Performer insgesamt nicht sehr unterschiedlich verhielten. Um weitere Erkenntnisse über die psychometrische Qualität des Performance Assessment via MMI zu gewinnen, muss die Stichprobe der Situationen und der Probanden erweitert werden. Ferner sollten die relevanten Verhaltensweisen auch nonverbale Facetten mit einbeziehen und Situationen gestaltet werden, in denen soziale Interaktionen mit anderen Personen durchzuführen sind.

■ Laufende Arbeiten

Um die Zahl der einsatzbereiten MMI-Situationen zu erhöhen und die inhaltliche Breite der Themen auszubauen, wurde im Studienjahr 2016/17 im Bachelorstudiengang Lehr-/Lern- und Trainingspsychologie ein Projektseminar angeboten, in dem die Studierenden ihre Bachelorarbeit anfertigen konnten. Im Rahmen des Seminars befassten sich die Studierenden

mit den Anforderungen des Lehrberufs und mit der Technik des Performance Assessments. Unter sehr intensiver Betreuung der beiden Autor*innen entwickelten die Studierenden fünf Situationen einschließlich des Materials für die Durchführung und Auswertung. Die Entwicklungsarbeit erwies sich als sehr aufwändig, sodass bezüglich der geplanten umfangreichen Erprobung der Situationen Abstriche gemacht werden mussten.

Tabelle 4 gibt einen Überblick über die im Zusammenhang mit dem Projektseminar erprobten Situationen. Dabei wurde die Situation 5a von den Studierenden in Eigenregie erprobt, während die anderen Situationen im Verbund durchgeführt wurden. Die Situationen mit dem Buchstaben „a" in der laufenden Nummer wurden von den Studierenden im Rahmen des Seminars entwickelt. Die Situationen mit dem Buchstaben „b" waren bereits in der Arbeit von Lüllemann und Simeth entwickelt worden. Hierzu entwickelte der Zweitautor Beurteilungsraster, während die Aufgaben weitgehend in der Originalversion verwendet wurden.

Tabelle 4
Im Rahmen des Projektseminars eingesetzten MMI-Situationen

	Situation	**Kompetenz**	**Typ**	**Kurzbeschreibung**
1a	Frustrierter Vater	Sachliche Konfliktlösung	Monolog	In einem Brief beklagt sich der Vater einer Schülerin über laufende Angriffe eines Mitschülers und droht Konsequenzen an. Dafür soll eine Problemlösung erarbeitet werden.
1b	Frustrierter Arzt	Sachliche Konfliktlösung	Monolog	Aus einem kritischen Forenbeitrag sollen die einzelnen, polemisch formulierten, Kritikpunkte aufgegriffen und jeweils mit einer sachlichen Stellungnahme versehen werden.
2a	Bullying	Empathie, Erziehungshandeln	Monolog	Ein lernschwacher Schüler wird zum Opfer eines Klassenkameraden. Die Situation soll analysiert und es soll eine Handlungsstrategie entwickelt werden.
2b	Moralisches Dilemma	Empathie, Erziehungshandeln	Monolog	Befreundete Eltern überlegen ihr Kind wegen befürchteter Hänseleien im Sportunterricht krank zu melden. Sie sollen als Lehrperson Rat geben, indem Sie Handlungsmöglichkeiten und deren Konsequenzen diskutieren.
3a	Unzuverlässiger Kollege	Konfliktfähigkeit, Problemlösung	Rollenspiel	Ein in der Zusammenarbeit unzuverlässiger Kollege soll zur Rede gestellt und eine Problemlösung erarbeitet werden.
3b	Projektwoche	Konfliktfähigkeit, Problemlösung	Rollenspiel	Eine uneinsichtige Lehrerkollegin hat sich für das bevorzugte und bereits vorbereitete Thema zur Projektwoche angemeldet, obwohl sie von dem Vorhaben der/des Bewerbenden wusste.
4a	Berufswunsch	Berufswahl, Selbstpräsentation	Monolog	Die eigenen Motive für die Berufswahl „Lehrer/in" sollen dargelegt, die persönliche Eignung beschrieben, Alternativen diskutiert werden
4b	Berufliche Herausforderungen	Berufswahl, Selbstpräsentation	Monolog	Bewerbende/r soll Herausforderungen im Lehrberuf, die eigene Motivation Lehrende/r zu werden und ihre/seine Eignung dafür beschreiben.
5a	Feedbackgespräch	Umgang mit Kritik	Interaktion	Ein relativ spontan gehaltener Kurzvortrag wird beurteilt; der/die Bewerbende soll zur vorgebrachten (nicht immer zutreffenden) Kritik Stellung nehmen

Als Beispiel soll die Aufgabenbeschreibung für Situation 3b wörtlich wiedergegeben werden:

> Nächste Woche ist Projektwoche. Sie haben sich schon wochenlang mit dem Thema „Mittelalter“ auseinander gesetzt und dieses intensiv vorbereitet. Sie freuen sich auf eine tolle Zeit mit kreativen und interessierten Schülerinnen und Schülern. Ihrer Kollegin, Frau Mayer, haben Sie ganz begeistert von Ihrem Vorhaben erzählt. Heute wollen Sie die Arbeit anmelden und stellen fest, dass Frau Mayer dasselbe Projekt soeben eingetragen hat. Darüber ärgern Sie sich sehr, denn laut Konferenzbeschluss soll jede Lehrperson ein separates Thema zur Verfügung stellen, um die Themenvielfalt zu gewährleisten. Sie machen sich sofort auf den Weg zu Frau Mayer in das Lehrerzimmer, um das Problem zu lösen.

In dieser Aufgabe sollten die Testpersonen zeigen, ob sie in der Lage sind, mit einem persönlichen Konflikt konstruktiv umzugehen, die eigenen Wünsche, aber auch die Situation des Gegenübers zu berücksichtigen und das Problem so zu lösen, dass eine weitere kollegiale Arbeit nicht beeinträchtigt sei.

Methode

Da die im Rahmen des Projektseminars gewonnenen Befunde an anderer Stelle ausführlicher dargestellt werden sollen, werden Durchführung, Auswertungsstrategien und Ergebnisse hier nur kurz präsentiert, zumal die Analysen zum Zeitpunkt der Berichtlegung noch nicht vollständig abgeschlossen sind.

Die Datenerhebung für die Situationen 1a bis 4b wurden von der Erstautorin in Form eines Stationenlaufs organisiert, wie dies bei der MMI-Durchführung üblich ist. Jeweils vier Kandidat*innen sollten gleichzeitig zur Testung antreten und in einem Rundlauf zunächst vier Situationen absolvieren. In einem weiteren Rundlauf, nach einer kurzen Pause, sollten sie die vier verbleibenden Stationen bestreiten. Alle Situationen waren so angelegt, dass die Aufgabenstellung innerhalb von 2-3 Minuten erfasst werden konnte und somit etwa 7 Minuten Zeit für die Ausführung blieben. Nach 10 Minuten fand jeweils der Wechsel zwischen den Stationen statt, sodass die Teilnehmer*innen genau 80 Minuten lang Aufgaben zu bearbeiten hatten.

Vorgesehen waren vier Testtermine mit jeweils vier Studierenden, die sich als Testkandidat*innen zur Verfügung gestellt hatten. Leider sprangen etliche Testpersonen relativ kurzfristig von der Teilnahme ab und es erwies sich als wenig aussichtsreich, weitere Durchführungen des Testparcours zu organisieren, da die Bereitstellung der Räume und der Aufzeichnungstechnik vom guten Willen vieler Beteiligter abhing.

Letztlich konnten innerhalb von zwei Tagen Daten von 13 Probanden erhoben werden. Sämtliche Darbietungen wurden aufgezeichnet, bei den Situationen 1a bis 3b audiovisuell, bei den Situationen 4a und 4b rein auditiv. Im Anschluss daran wurden die Aufzeichnungen technisch bearbeitet und einige wenige Darbietungen als unbrauchbar ausgeschlossen. Dies war dann der Fall, wenn sich der Eindruck aufdrängte, dass die Probanden die Situation nicht ernsthaft bearbeitet hatten oder von der Aufgabenstellung schlicht überfordert waren. Die verbleibenden Aufzeichnungen wurden von den Studierenden des Projektseminars mithilfe der erstellten Beurteilungsbögen ausgewertet. Die Beurteilungsbögen wurden daraufhin den Studierenden übergeben, die die jeweilige Situation entwickelt hatten, und schließlich von ihnen statistisch ausgewertet.

Ergebnis

Um einen Einblick in die Auswertungen zu vermitteln, sollen beispielhaft die Reliabilitätsanalysen für die Situation 3b dargestellt werden. In dieser Situation traf die Testperson im Rollenspiel auf eine Kollegin (eine Psychologiestudentin, die genaue Verhaltensanweisungen erhalten hatte), die ihr das Thema zu einer schulischen Projektwoche „weggeschnappt" hatte. Die Testperson sollte nun klären, wie es zu diesem Verhalten gekommen sei, und sollte eine Konfliktlösung entwickeln. Die meisten Testpersonen gingen unerwartet freundlich mit der eigentlich unfair agierenden Kollegin um und boten ihr im Regelfall an, das Thema für die Projektwoche gemeinsam zu gestalten.

Die Performance der Testpersonen wurde mit einem verhaltensverankerten Beurteilungsraster aufgezeichnet, das neun jeweils vierstufige Skalen enthielt. Die inhaltlichen Dimensionen für die Beurteilung wurden zunächst deduktiv aus Modellen der Konfliktbearbeitung abgeleitet. Dementsprechend sollte die handelnde Person die eigene Betroffenheit deutlich machen, versuchen, das Verhalten der anderen Person zu verstehen, eigene Wünsche und Ziele formulieren, Lösungsvorschläge erarbeiten und ggf. ausdiskutieren sowie versuchen, zu einer gütlichen Einigung zu kommen. Nach einem ersten Entwurf der Beurteilungsdimensionen wurden einige Videoaufzeichnungen in Augenschein genommen, um zu prüfen, ob die erwarteten Verhaltensweisen auftraten und beobachtet werden konnten. Dies führte zu kleineren Änderungen in der Liste der Beurteilungsdimensionen, vor allem aber zur konkreten Ausformulierung der Beurteilungsstufen. Für jede Dimension sollten vier Niveaustufen formuliert werden, die grundsätzlich diesen Kriterien folgen sollten:

- (0) liegt weitab vom erwarteten Standard,
- (1) nähert sich dem erwarteten Standard an,
- (2) erfüllt den erwarteten Standard,
- (3) liegt über dem erwarteten Standard.

Aufgrund der Analyse einiger Videoaufzeichnungen wurden diese vier Stufen am Beispiel der Dimension „Äußerung der eigenen Betroffenheit und Frustration" wie folgt ausformuliert:

- (0) Zeigt keine erkennbare Betroffenheit <u>oder</u> äußert die Betroffenheit extrem stark.
- (1) Zeigt etwas Betroffenheit <u>oder</u> übertreibt die Betroffenheit etwas.
- (2) Spricht die Betroffenheit und Frustration in angemessener Gefühlslage an.
- (3) Spricht die Betroffenheit und Frustration in angemessener Gefühlslage an <u>und</u> äußert Irritation hinsichtlich der weiteren kollegialen Zusammenarbeit.

Bei den Qualitätsstufen (0) und (1) sind jeweils zwei alternative Verhaltensweisen (mit „oder" getrennt) angegeben, die zwar zu derselben numerischen Einschätzung führen, bei der individuellen Rückmeldung jedoch qualitativ unterschieden werden können.

Für die Berechnung der ICC-Koeffizienten zur Abschätzung der Reliabilität nach der weiter oben beschriebenen Methode wurden die Daten von acht Performern und sechs Ratern verwendet, da diese einen vollständigen Datensatz ergaben.

Berechnet wurden die Reliabilitäten für einen Rater [ICC(2,1)] und für den Durchschnitt aus den verfügbaren sechs Ratern [ICC(2,6)]. Um abzuschätzen, welchen Effekt es hätte, wenn drei Rater eingesetzt würden, wurden zufällig zwei Gruppen aus je drei Ratern gebildet und dafür die ICC-Werte berechnet [ICC(2,3)a und ICC(2,3)b].

Wie Tabelle 5 zeigt, fallen die Reliabilitätskoeffizienten relativ hoch aus, wenn berücksichtigt wird, dass die Einschätzung konkreter Verhaltensweisen immer schwierig ist. Wie zu erwarten, sind die Beurteilungen durch eine einzelne Person nur von mittlerer Zuverlässigkeit; betrachtet man jedoch den Durchschnitt von sechs Ratern, so liegen fast alle Koeffizienten bei oder über dem Wert von 0.8. Probehalber wurden zwei zufällige Gruppen von jeweils drei Ratern gebildet und hierfür mittlere Beurteilungen berechnet. In etlichen Fällen zeigt sich, dass es bereits bei der Verfügbarkeit von drei Ratern möglich ist, sehr zuverlässige Beurteilungen zu erhalten. Allerdings scheint nicht jede Beurteilergruppe zu eindeutigen Einschätzungen zu gelangen, sodass die Auswahl bzw. das Training der Rater ebenfalls eine wichtige Rolle für die Präzision der Ergebnisse spielen dürfte.

Tabelle 5: *ICC-Koeffizienten für die Beurteilungsdimensionen der Konfliktlösung mit einer Arbeitskollegin*

		ICC(2,1)	**ICC(2,3)a**	**ICC(2,3)b**	**ICC(2,6)**
1.	Äußerung der eigenen Betroffenheit und Frustration	.40	.65	.65	.80
2.	Selbstbeschreibung der Situationswahrnehmung (habe das Thema ausgewählt und vorbereitet, der Kollegin davon erzählt, angenommen, dass das Thema respektiert wird)	.57	.81	.75	.89
3.	Verdeutlichung der eigenen Situation (Arbeitsaufwand)	.30	.38	.72	.72
4.	Aufklärung des Verhaltens des Gegenübers	.54	.73	.86	.88
5.	Verdeutlichung der Erwartungen an das Gegenüber (Kompromissbereitschaft, Einhalten von Regeln)	.53	.85	.54	.87
6.	Vorlegen von Lösungsvorschlägen	.48	.91	.41	.85
7.	Finden einer akzeptablen, die eigene Person nicht benachteiligenden Lösung und Widerstand gegen weitere Benachteiligung	.80	.92	.97	.96
8.	Akzeptanz der Lösung	.77	.93	.91	.95
9.	Sichern einer kollegialen Arbeitsbeziehung	.73	.89	.89	.94
	Gesamtsumme	.75	.88	.93	.95

Die vorgegebene Situation erwies sich als durchaus anspruchsvoll und die Verhaltensweisen der Testpersonen als nicht unproblematisch. Kaum eine Person versuchte wirklich zu klären, warum die Kollegin das Thema der Projektwoche „plagiiert“ hatte, und kaum eine Person versuchte die Kollegin darauf zu verpflichten, künftig offen und ehrlich zu agieren. Allerdings hatte die schauspielende Kollegin relativ enge Anweisungen erhalten, sodass ein ausgiebiges und vielschichtiges Gespräch nicht zustande kam. Als sehr positiv ist jedoch der verbindliche und konstruktive Ansatz vieler Testpersonen zu sehen, über das unangemessene Verhalten der Kollegin hinwegzusehen und eine kooperative Lösung zu erwirken. Da einzelne Testper-

sonen jedoch auch distanziert agierten, zeigten sich deutliche Differenzen zwischen den Probanden, die ein wesentlicher Grund für die beobachteten hohen Reliabilitäten darstellen. Insgesamt erwies sich die Situation als interessante Gelegenheit, soziale Umgangsformen unter kritischen Bedingungen zu beobachten. Es ist zu vermuten, dass die Teilnehmenden eine differenzierte Rückmeldung zu ihrem Verhalten und den Besonderheiten ihres Verhaltens im Vergleich zu den anderen Testpersonen als sehr informativ beurteilen würden. Ebenfalls ist davon auszugehen, dass die verwendete Rollenspielsituation einen guten Ausgangspunkt für die Reflexion über den eigenen Umgang mit Konflikten darstellt und einen motivierenden Ansatz für ein Konfliktlösetraining bietet.

Erkenntnisstand

Insgesamt wurden im Frühjahr und Sommer 2017 neun MMI-Situationen und die dazu konstruierten Beurteilungsinstrumente einer Erprobung unterzogen. Auch wenn noch nicht alle Situationen und Daten ausgewertet sind, lässt sich bereits jetzt sagen, dass die wesentlichen Testgütekriterien (Objektivität, Reliabilität und Validität) durch überlegte Konstruktion der situativen Anforderungen und der Auswertungsinstrumente sowie durch ein gutes Training der Beurteilenden gewährleistet werden. Es scheint allerdings unabdingbar, dass jede Verhaltensleistung unabhängig von mehreren Beurteilenden eingeschätzt wird und dass die einzelnen Bewertungen zu einer Gesamteinschätzung auf den jeweiligen Dimensionen kombiniert werden. Eine sehr zuverlässige Beurteilung der individuellen Verhaltensleistung scheint durch Verrechnung der einzelnen Dimensionen möglich zu sein. Dadurch ergibt sich pro Situation eine ausreichend differenzierte Skala, die Veränderungen als Folge von Trainingsmaßnahmen sensibel abbilden, die sich aber auch für Selektionszwecke eignen dürfte.

Eine besondere Schwierigkeit war, Studierende zur freiwilligen Teilnahme am Performance Assessment zu motivieren, selbst bei Vorgabe eines finanziellen Anreizes. Die Scheu, das eigene Verhalten in einer Art Testsituation zu präsentieren und dieses aufzeichnen zu lassen, erwies sich als relativ stark ausgeprägt. Allerdings äußerten die Teilnehmenden nach der Durchführung der Aufgaben, dass die Anforderungen durchaus zu bewältigen waren und die Situationen interessante Einblicke in das eigene Verhalten ermöglichten. Die hohe Akzeptanz des verhaltensdiagnostischen Ansatzes aufgrund der erlebten Authentizität dürfte eine gute Motivation für anschließende Trainingsangebote liefern.

Bei der Erhebung der Daten war eine Situation angestrebt worden, wie sie für die Auswahlverfahren an medizinischen Einrichtungen typisch sind: Die Stationen waren der Reihe nach ohne größere Pausen, ohne intensivere Anleitungen und ohne die Möglichkeit des Austausches untereinander zu durchlaufen. Diese Durchführung erwies sich als durchaus anstrengend für alle Beteiligten und insichtlich der Organisation der Personen, der Räume, des Equipments und der Abläufe als sehr aufwändig. Insgesamt muss der hohe Aufwand für Durchführung, Aufzeichnung und Auswertung noch optimiert werden. Wie an der Universität Wien (https://www.schroedingerskatze.at/doktorspiele/) müsste eine Gruppe von Schauspielenden aufgebaut werden, die für Assessments regelmäßig zur Verfügung stünde. Ein ständig verfügbares Aufnahmestudio mit der erforderlichen Technik bzw. den Requisiten sowie Personal für die Durchführung würden den Organisationsaufwand deutlich verringern.

Bei der Gestaltung der Situationen im Hinblick auf die zu erfassenden Kompetenzen wurden Ideen aus den medizinischen Zulassungsverfahren verwendet und einzelne Kompetenzen

eher isoliert betrachtet. Es fehlt jedoch noch ein systematischer Katalog von Verhaltensdispositionen, die für den Lehrberuf inhaltlich wirklich relevant und per Performance Assessment besonders gut erfassbar sind. In diese Richtung werden die künftigen Arbeiten gehen. Dabei wird es vor allem erforderlich sein, intuitive Konzepte und Strategien bei der Planung und Durchführung von Unterricht mithilfe passender Situationen zu ermitteln. Wie bereits dargelegt, sind es nicht nur Persönlichkeitsmerkmale, die für die erfolgreiche Bewältigung des Lehramtsstudiums und des Lehrberufs bedeutsam sind. Es sind gerade auch subjektive Konzeptionen des Lehrens und Lernens, sogenannte „Theorien mittlerer Reichweite“ (Gottein, 2016), die das persönliche Unterrichtshandeln steuern und die nur schwer modifizierbar sind. Eine frühe diagnostische Erfassung solcher pädagogischer Fehlkonzepte wäre für eine intensive Bearbeitung im Rahmen des Studiums, vor allem in den schulpraktischen Ausbildungseinheiten, von großer Bedeutung.

Literatur

Abdul Rahim, A. F. & Yusoff, M. S. B. (2016). Validity evidence of a Multiple Mini Interview for selection of medical students: Universiti Sains Malaysia Experience. *Education in Medicine Journal, 8*(2), 49-63.

Brownell, K., Lockyer, J., Collin, T. & Lemay, J. F. (2007). Introduction of the multiple mini interview into the admissions process at the University of Calgary: acceptability and feasibility. *Medical Teacher, 29*(4), 394-396.

Callwood, A. (2015). *Developing and evaluating the Multiple Mini Interview in student midwife selection. Doctoral Thesis.* Guildford: University of Surrey, Faculty of Health and Medical Sciences.

Campagna-Vaillancourt, M., Manoukian, J., Razack, S., Nguyen, L.H. (2014). Acceptability and reliability of multiple mini interviews for admission to otolaryngology residency. *Laryngoscope, 124*, 91-96.

Cleland, J. A., Dowell, J., McLachlan, J., Nicholson, S. & Patterson, F. (2012). Identifying best practice in the selection of medical students: literature review and interview survey. *General Medical Council.* (Online verfügbar unter http://www.gmc-uk.org/Identifying_best_practice_in_the_selection_of_medical_students.pdf_51119804.pdf; letzter Abruf: 31. 8. 2017).

Cox, W. C., McLaughlin, J. E., Singer, D., Lewis, M. & Dinkins, M. M. (2015). Development and assessment of the multiple mini-interview in a school of pharmacy admissions model. *American Journal of Pharmaceutical Education, 79*(4), Article 53.

Dowell, J., Lynch, B., Till, H., Kumwenda, B. & Husbands, A. (2012). The multiple mini-interview in the UK context: 3 years of experience at Dundee. *Medical Teacher, 34*(4), 297–304.

Eva, K. W., Rosenfeld, J., Reiter, H. I. & Norman, G. R. (2004). An admissions OSCE: The multiple mini-interview. *Medical Education, 38*(3), 314-326.

Frey, A. (2008). *Kompetenzstrukturen von Studierenden in der ersten und zweiten Phase der Lehrerbildung. Eine nationale und internationale Standortbestimmung.* Landau: Empirische Pädagogik.

Frost, F. M. M. (2015). *Auswahlgespräche als Instrument der Eignungsprüfung zum Lehramtsstudium. Möglichkeiten und Grenzen am Beispiel der TUM School of Education. Dissertation.* München: Technische Universität, TUM School of Education.

Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D. & Ziv, A. (2012). Reliability estimates: Behavioural stations and questionnaires in medical school admissions. *Medical Education, 46(3),* 277-288.

Gottein, H.-P. (2016). *Tun sie denn, was sie wissen? Hochschuldidaktische Überlegungen für eine kompetenzorientierte und handlungspsychologisch begründete Lernumgebung in der Ausbildung von Lehrerinnen und Lehrern*. Bad Heilbrunn: Klinkhardt.

Harasym, P. H., Woloschuk, W., Mandin, H. & Brundin-Mather, R. (1996). Reliability and validity of interviewers' judgments of medical school candidates. *Academic Medicine: Journal of the Association of American Medical Colleges, 71(1 Suppl.)*, 40-42.

Heldenbrand, S. D., Flowers, S. K., Bordelon, B. J., Gubbins, P. O., O'Brien, C., Stowe, C. D. & Martin, B. C. (2016). Multiple mini-interview performance predicts academic difficulty in the PharmD curriculum. *American Journal of Pharmaceutical Education, 80*(2), Article 27.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.

Hines, C., Cruickshank, D. & Kennedy, J. (1985). Teacher clarity and its relationship to student achievement and satisfaction. *American Educational Research Journal, 22*(1), 87-99.

Hissbach, J. C., Sehner, S., Harendza, S. & Hampe, W. (2014). Cutting costs of multiple mini-interviews – changes in reliability and efficiency of the Hamburg medical school admission test between two applications. *BMC Medical Education*, *14*, 54-63.

Ho, A. D. & Kane, T. J. (2013). *The reliability of classroom observations. MET Project Research Paper.* Seattle: Bill & Melinda Gates Foundation.

Humphrey, S., Dowson, S., Wall, D., Diwakar, V. & Goodyear, H. M. (2008). Multiple mini-interviews: Opinions of candidates and interviewers. *Medical Education, 42*, 207-213.

Husbands, A. & Dowell, J. (2013). Predictive validity of the Dundee multiple mini-interview. *Medical Education*, *47*(7), 717-725.

Kelly, M. E., Dowell, J., Husbands, A., Kropmans, T., Jackson, A. E., Dunne, F., O'Flynn, S., Newell, J., Murphy, A.W. (2014a). Can multiple mini interviews work in an Irish setting? A feasibility study. *Irish Medical Journal, 107*, 210-212.

Kelly, M. E., Dowell, J., Husbands, A., Newell, J., O'Flynn, S., Kropmans, T., Dunne, F. P., Murphy, A.W. (2014b). The fairness, predictive validity and acceptability of multiple mini interview in an internationally diverse student population – a mixed methods study. *BMC Medical Education*, *14*, 267-279.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Expertise.* Bonn: BMBF.

KMK. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften*. Beschluss der Kultusministerkonferenz vom 16.12.2004.

Knorr, M. & Hissbach, J. (2014). Multiple mini-interviews: Same concept, different approaches. *Medical Education*, *48*(12), 1157-1175.

Košinár, J. (2014). *Professionalisierungsverläufe in der Lehrerausbildung: Anforderungsbearbeitung und Kompetenzentwicklung im Referendariat.* Berlin: Barbara Budrich.

Krapp, A. (1979). *Prognose und Entscheidung.* Weinheim: Beltz.

Laurence, C. O., Zajac, I. T., Lorimer, M., Turnbull, D. A. & Sumner, K. E. (2013). The impact of preparatory activities on medical school selection outcomes: A cross-sectional survey of applicants to the university of Adelaide medical school in 2007. *BMC Medical Education, 13*(1), 159-167.

Lüllemann, J. & Simeth, N. (2016). *Multiple Mini-Interviews als Zulassungsverfahren für das Lehramtsstudium – Entwicklung eines eignungsdiagnostischen Verfahrens. Projektarbeit.* Erfurt: Universität Erfurt, Fachgebiet Psychologie.

Lüllemann, J. (2016). *Eignungsdiagnostik für das Lehramt – Eine Erfassung der Interrater-Reliabilität der Multiplen Mini-Interviews. Masterarbeit.* Erfurt: Universität Erfurt, Fachgebiet Psychologie.

Messick, S. (1995). Validity of psychological assessment: Validation of references from persons' responses and performances on scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Moede, W. (1930). *Lehrbuch der Psychotechnik*. Berlin: Springer.

Nolle, T. (2016). Eignungsvoraussetzungen für einen sich ständig verändernden Beruf. In A. Boeger (Hrsg.), *Eignung für den Lehrerberuf: Auswahl und Förderung* (S. 13-30). Wiesbaden: Springer.

Oliver, T., Hecker, K., Hausdorf, P. A. & Conlon, P. (2014). Validating MMI scores: are we measuring multiple attributes? A*dvances in Health Sciences Education, 19*(3), 379–392.

Palm, T. (2001). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation, 39,* 1-11.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F. & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*(1), 36-60.

Pau, A., Jeevaratnam, K., Chen, Y. S., Fall, A. A., Khoo, C. & Nadarajah, V. D. (2013). The Multiple Mini-Interview (MMI) for student selection in health professions training - A systematic review. *Medical Teacher, 35*, 1027-1041.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2-12.

Razack, S., Faremo, S., Drolet, F., Snell, L., Wiseman, J., Pickering, J. (2009). Multiple mini-interviews versus traditional interviews: Stakeholder acceptability comparison. *Medical Education, 43*(10), 992-1000.

Rees, E. L., Hawarden, A. W., Dent, G., Hays, R., Bates, J. & Hassell, A. B. (2016). Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review: BEME Guide No. 37. *Medical Teacher, 38*(5), 443-55.

Ribera, T., BrckaLorenz, A., Cole, E. R. & Laird, T. F. L. (2012). Examining the importance of teaching clarity: Findings form the Faculty Survey of Student Engagement. *Annual Meeting of the American Educational Research Association*, April 13-17, 2012, Vancouver, BC, Canada.

Roberts, C., Clark, T., Burgess, A., Frommer, M., Grant, M. & Mossman, K. (2014). The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. *BMC Medical Education, 14,* 169-179.

Roberts, C., Rothnie, I., Zoanetti, N. & Crossley, J. (2010). Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical Education, 44*(7), 690–698.

Rothland, M. & Terhart, E. (2011). Eignungsabklärung angehender Lehrerinnen und Lehrer. Einführung in den Thementeil. *Zeitschrift für Pädagogik, 57*(5), 635-638.

Rothland, M. (2013). Allgemeine Persönlichkeitsmerkmale als Eignungskriterien für den Lehrerberuf? Eine Folgestudie. *Lehrerbildung auf dem Prüfstand, 6*(1), 70-91.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schuler, H. (2007). Spielwiese für Laien? Weshalb das Assessment-Center seinem Ruf nicht mehr gerecht wird. *Wirtschaftspsychologie aktuell, 2/2007,* 27-30.

Sebok, S. S., Luu, K. & Klinger, D. A. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: Findings from generalizability and Rasch analyses. *Advances in Health Sciences Education*, *19*(1), 71-84.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Simmenroth-Nayda, A., Meskauskas, E., Burckhardt, G. & Görlich, Y. (2014). Das neue Göttinger Auswahlverfahren für Medizin - welche Bewerber können profitieren? *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 108*(10), 609–617.

Simonds, C. J. (1997). Classroom understanding: An expanded notion of teacher clarity. *Communication Research Reports, 14*(3), 279-290.

Stemmler, G. & Margraf-Stiksrud, J. (Hrsg.). (2015). *Lehrbuch psychologische Diagnostik.* Bern: Huber.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement, 6,* 33-42.

Titsworth, S., Mazer, J. P., Goodboy, A. K., Bolkan, S. & Myers S. A. (2015). Two meta-analyses exploring the relationship between teacher clarity and student learning. *Communication Education, 64*(4), 385-418.

Uijtdehaage, S., Doyle, L. H. & Parker, N. (2011). Enhancing the reliability of the multiple mini-interview for selecting prospective health care leaders. *Academic Medicine*, *86*(8), 1032-1039.