Christoph J. Kemper, Tina Mitschke,
Wolfram Rollett, Viktoria Kemper & Daniel Oberfeld

Verbesserung der Präsentationskompetenz in der Lehre an deutschen Hochschulen

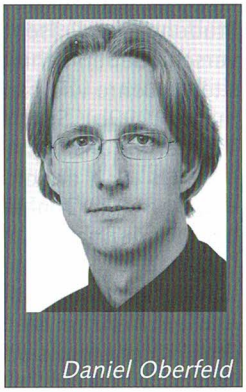
Entwicklung des Mainzer Verfahrens zur Peer-Evaluation studentischer Präsentationen (MPEP)

Christoph J. Kemper

Tina Mitschke

Wolfram Rollett

Viktoria Kemper


Daniel Oberfeld

Good presenting skills is a critical competence for many employees. When studying at the university, students usually have numerous opportunities to develop these skills as long as appropriate feedback is provided for their presentation performance. In the research presented, a questionnaire for the systematic evaluation of student presentation skills by their peers was developed and validated (reliability, factorial validity, criterion-related validity). The aim of the study was to provide a useful assessment instrument for the systematic evaluation and structured feedback of student presentation skills to teachers and students, thereby contributing to the quality of university teaching.

Das fachlich korrekte und ansprechende, mediengestützte Präsentieren ist eine erfolgskritische Kompetenz für viele Erwerbstätige. Während eines Hochschulstudiums bieten sich zahlreiche Gelegenheiten, diese Kompetenz zu entwickeln, sofern den Studierenden Präsentationsleistungen adäquat zurückgemeldet werden. In der hier vorgestellten dreiphasigen Untersuchung wurde ein Fragebogen für die systematische Bewertung studentischer Präsentationsleistungen seitens der Veranstaltungsteilnehmer (Peer-Evaluation) entwickelt und Güte, Akzeptanz und Nutzen überprüft. In der Hochschulpraxis erfolgt die Rückmeldung studentischer Präsentationsleistungen bisweilen unstrukturiert und unsystematisch. Die wenigen verfügbaren Fragebögen wurden bisher nicht evaluiert. Ziel der Untersuchung war es daher, Lehrenden und Lernenden an deutschen Hochschulen ein zeiteffizientes, geprüftes und nützliches Instrument für die systematische Evaluation und strukturierte Rückmeldung studentischer Präsentationsleistungen zur Verfügung zu stellen und somit einen Beitrag zur Qualitätssicherung in der Lehre zu leisten.

1. Einleitung

Das mediengestützte Präsentieren ist eine Anforderung, die aus dem heutigen Arbeitsalltag vieler Menschen nicht mehr wegzudenken ist (Wippermann 2014). Die Vermittlung entsprechender Kompetenzen, die für das erfolgreiche Präsentieren benötigt werden, ist daher ein wichtiger Bestandteil der pädagogischen Arbeit an Schulen und Hochschulen. Präsentationskompetenz bereits in den Hochschulen, also vor dem Eintritt in das Erwerbsleben zu vermitteln, stellt daher ein erstrebenswertes bildungspolitisches Ziel dar. Damit die Studierenden, die während ihres Studiums ohnehin zahlreiche Präsentationsleistungen erbringen müssen, ihre Präsentationskompetenz verbessern können, ist eine konstruktive und informative Rückmeldung der erbrachten Leistung notwendig (Frese/Zapf 1994; Gibbs/Simpson 2004; Nicol/Macfarlane-Dick 2006). Üblicherweise werden Präsentationsleistungen von den Veranstaltungsleitern zurückgemeldet. Eine weitere nützliche Informationsquelle für die Rückmeldung der Leistung der Präsentie-

renden bleibt allerdings oft ungenutzt: die übrigen Teilnehmer einer Lehrveranstaltung, die Mitstudierenden (fortan auch: Peers). Im vorliegenden Beitrag wird die Entwicklung und Validierung eines diagnostischen Verfahrens für die Peer-Evaluation studentischer Präsentationen geschildert, das eine systematische und informative Rückmeldung von Präsentationsleistungen ermöglicht. Die empirische Grundlage des Mainzer Verfahrens zur Peer-Evaluation studentischer Präsentationen (MPEP) bildet eine Serie von empirischen Erhebungen, in denen studentische Präsentationen durch Peers bewertet wurden. Im Folgenden wird zunächst der theoretische Hintergrund der Arbeit geschildert, bevor auf die empirischen Erhebungen eingegangen wird.

Studentische Präsentationen bilden das Fundament von verschiedenen Lehrveranstaltungsformaten in der universitären Lehre wie bspw. Seminaren und Tutorien (Rindermann 1999b). Sie beanspruchen oft einen wesentlichen Teil der Veranstaltungszeit. Daher trägt sowohl die fachliche als auch die didaktische Qualität der studentischen Beiträge zur Qualität der gesamten Lehrveranstaltung bei. Rindermann (1999) konnte z.B. zeigen, dass die Qualität studentischer Referate ein relevanter Prädiktor von Seminarevaluationen ist. Studentische Präsentationen erfüllen für die Studierenden wichtige Funktionen. Ihre Präsentationsleistungen demonstrieren, dass sie sich innerhalb kurzer Zeit aktiv in ein neues Themengebiet einarbeiten können und in der Lage sind, den Lehrstoff soweit zu durchdringen und aufzubereiten, dass sie diesen den Mitstudierenden vermitteln können. Zudem bietet die Präsentation des Lehrstoffs vor dem Dozenten[1] und den Peers eine wichtige Lerngelegenheit zum Erwerb und zur Verbesserung der Präsentationskompetenz (vgl. Bromme/Rambow 1993).

Das Präsentieren ist eine komplexe Leistung, die mit mehreren Teilkompetenzen wie den Fachkompetenzen (z.B. Sachverhalte wissen, Argumente verstehen und bewerten, Medien fachspezifisch einsetzen), Selbstkompetenzen (z.B. Arbeitszeit planen, selbstständig und verantwortlich handeln), lernmethodischen Kompetenzen (z.B. zielgerichtet recherchieren, Inhalte strukturieren und darstellen) und Sozialkompetenzen (z.B. adressatengerecht vortragen, kritikfähig sein) in Verbindung gebracht wird (Wippermann 2014). Die Bedeutung der Präsentationskompetenz hat in den vergangenen Jahren deutlich zugenommen. Schnack (2011) führt diesen Trend zum Einen auf eine durch den technischen Fortschritt bedingte Präferenz für audiovisuelle Darbietungen zurück. Zum Anderen beinhalte die mediengestützte Präsentation Anforderungen, mit denen sich Menschen heute in Schule, Studium, Ausbildung oder Beruf auseinandersetzen müssen: selbständiges Lernen und Arbeiten, den Umgang mit audiovisuellen Medien, die Inszenierung von Inhalten und das freie Sprechen. Die mediengestützte Präsentation sei inzwischen ein weit verbreiteter Modus zur Vermittlung von Arbeitsergebnissen. Verschiedene Bundesländer haben auf diesen Trend reagiert und die mediengestützte Präsentation bspw. als festen Bestandteil des schulischen Unterrichts verankert (Schnack 2011).

Eine notwendige Voraussetzung für die Verbesserung von Präsentationskompetenz ist eine diagnostisch gehaltvolle Rückmeldung über die erbrachte Leistung. Qualifiziertes Feedback gilt als eine der besten Möglichkeiten, um Ansatzpunkte zur Verbesserung der eigenen Leistung zu finden (Frese/Zapf 1994; Gibbs/Simpson 2004; Nicol/Macfarlane-Dick 2006). Die Feedback-Kultur im Hochschulalltag, die noch vor einigen Jahren wenig ausgeprägt war (Gold/Souvignier 2000), ist inzwischen in den meisten Ländern gesetzlich verankert und an den Hochschulen institutionalisiert (Rindermann 2003).

Die systematische Evaluation und strukturierte Rückmeldung studentischer Präsentationsleistungen durch Dozenten ist bisher allerdings noch immer eher die Ausnahme als die Regel. Wenn Präsentationsleistungen an Studierende zurückgemeldet werden, geschieht dies meist mündlich, summarisch und nicht immer in strukturierter Form. Hinzu kommt, dass eine weitere Quelle nützlicher Informationen bisher weitgehend ungenutzt bleibt – die Adressaten der Präsentation. Neben der Rückmeldung des Dozenten ist für viele Studierende auch das Urteil ihrer Mitstudierenden, die Peer-Evaluation interessant, denn *„inwieweit das Referat die Adressaten erreicht hat, können nur diese selbst beurteilen und rückmelden"* (Preiser 1995, S. 115). Mit dem Begriff Peer-Evaluation ist generell die Bewertung von studentischen Leistungen durch die Kommilitonen, die „Peers" gemeint. Topping (2009) definiert diese Bewertung als *„arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners"* (p. 20). Als „products" können nach Topping (2009) z.B. schriftliche Arbeiten, Präsentationen, Testleistungen oder Fertigkeiten angesehen werden. Im Folgenden wird auf einen bestimmten Bereich von Peer-Evaluation fokussiert: die Evaluation von studentischen Präsentationen.

Bisherige Forschungsergebnisse zeigen, dass mithilfe von Peer-Evaluationen Präsentationsleistungen nicht nur valide beschrieben werden können (Übereinstimmung von Peer- und Dozenten-Ratings: $r = .66$; Metaanalyse von Falchikov/Goldfinch 2000), sondern diese auch eine hohe Akzeptanz bei allen beteiligten Personen genießen und ihr Nutzen für die Verbesserung der Präsentationskompetenz durch verschiedene Studien nahe gelegt wird (Falchikov/Goldfinch 2000; Gold/Souvignier 2000; Saito/Fujita 2004; Topping 1998). Topping (1998) beschreibt in seinem Übersichtsartikel, dass die Vorteile der Peer-Evaluation für die Bewerteten darin bestehen, klarer identifizieren zu können, welche Aspekte zu einer hohen Qualität der eigenen Arbeit führen. Gold/Souvignier (2000) erfassten in ihrer Studie die Reaktionen von Studierenden auf die Bewertung durch Mitstudierende und Dozenten. Bewertet wurden die Referate der Studierenden anhand eines Fragebogens. Die Bewertungen wurden in Gesprächen mit studentischen Tutoren rückgemeldet und mittlere Beurteilungen in Graphiken und Tabellen sowie die häufigsten freien Antworten dargestellt. Fast alle Referenten bezeichneten die Ergebnisrückmeldung als verständlich und angemessen. Auch in

[1] Nach Möglichkeit verwenden wir gender-neutrale Bezeichnungen. Wo dies nicht sinnvoll umsetzbar ist, werden keine unterschiedlichen Genusformen verwendet, um künstliche Wortkombinationen zu vermeiden. Wir verwenden in diesen Fällen die männliche Genusform, die generisch gemeint ist.

Bezug auf die Inhalte des Feedbacks fanden Gold/Souvignier (2000) eine hohe Akzeptanz: 84% der Referenten bezeichneten die Urteile der Peers als nachvollziehbar und fair. Besondere Vorteile sahen 81% darin, dass durch das Feedback Ansatzpunkte für Verhaltensänderungen identifiziert werden.

Damit das Feedback zu Verhaltensänderungen und Kompetenzzuwächsen führen kann, sind nach Gibbs/Simpson (2004) folgende Bedingungen förderlich: Das Feedback sollte z.B. verhaltensnah und ausreichend häufig sein, sollte auf das Lernen und die Leistung des Lernenden und nicht auf Personenmerkmale fokussiert sein, sollte zum passenden Zeitpunkt erfolgen, sollte dem Ziel der geleisteten Aufgabe angemessen sein und muss rezipiert und in Verhalten umgesetzt werden. Topping (1998) berichtet in seiner Übersichtsarbeit verschiedene Studien zu den Effekten von Peer-Feedback und führt an: *"There is substantial evidence that peer assessment can result in improvements in the effectiveness and quality of learning ..."* (p. 22).

Verfahren zur Messung studentischer Präsentationsleistungen in der universitären Lehre sind bisher rar. Falchikov (1995) entwickelte gemeinsam mit Studierenden vier Kriterien anhand derer zehnminütige Präsentationen zu Fachaufsätzen aus der Entwicklungspsychologie bewertet wurden. Die Kriterien *structure, knowledge of topic, amount of information* und *delivery* wurden durch Sätze näher beschrieben. Weiterhin wurden die besten und schlechtesten Aspekte der Präsentation notiert sowie eine Gesamtnote vergeben. Ähnlich ging MacAlpine (1999) bei der Bewertungen von Präsentationen im Seminar High Voltage Engineering vor. Die Kriterien *knowledge, voice, body language* und *overall effect* wurden auf einer fünfstufigen Ratingskala eingeschätzt. Kurze Beschreibungen der Kriterien waren im Fragebogen enthalten, um diese zu präzisieren. Auch bei Langan/Shukur/Cullen/Penney/Preziosi/Wheater (2008) wurden wenige abstrakte Kriterien *(presentation, content* und *structure)* als Grundlage der Bewertung verwendet, bevor diese in unterschiedlichen Gewichtungen zu einer Gesamtnote integriert wurden. Den drei genannten Verfahren ist gemeinsam, dass sie zwar ökonomisch sind aber Einschränkungen im Hinblick auf Nutzen und Validität der damit möglichen Peer-Evaluation haben. Die verwendeten Bewertungskriterien und die zum Teil sehr allgemein gehaltenen Beschreibungen sind relativ abstrakt. Es ist daher fraglich, ob ein beurteilerübergreifendes Verständnis der Bewertungsdimensionen ohne Training der Beurteiler gewährleistet werden kann und ob die Beurteilten vor diesem Hintergrund überhaupt von der Rückmeldung profitieren. Gold/Souvignier (2000) stellen hingegen einen Fragebogen zur Bewertung studentischer Referate (FSR) vor, der 19 verhaltensnahe Items enthält, die einen höheren Nutzen im Hinblick auf die Rückmeldung haben. Leider berichten die Autoren keine psychometrischen Kennwerte für ihren Fragebogen.

Ziel der vorliegenden Untersuchung war es, ein neues Instrument zur Peer-Evaluation zu entwickeln, das für den standardmäßigen Einsatz in universitären Lehrveranstaltungen geeignet ist. Die Kriterien, die das neu zu entwickelnde Verfahren erfüllen sollte, ergaben sich z.T. aus dieser Zielsetzung. Um Dozenten zu überzeugen, das Verfahren in ihrer Lehre einzusetzen, musste es kurz und zeitökonomisch sein, damit die Bewertungen nicht zu viel Veranstaltungszeit in Anspruch nehmen. Um die Studierenden zu überzeugen, mussten Nutzen- und Akzeptanzaspekte im Vordergrund stehen. Daher wurden die Studierenden am Konstruktionsprozess beteiligt. Zudem wurden, im Gegensatz zu vielen früheren Verfahren, verhaltensnahe Indikatoren (Testitems) mit höherem diagnostischem Nutzen für die Rückmeldung entwickelt (vgl. Gibbs/Simpson 2004). Allerdings wurde auch darauf geachtet, dass die Indikatoren nicht zu spezifisch werden, damit das Verfahren flexibel über verschiedene Fächer bzw. Fachbereiche hinweg eingesetzt werden kann. Schließlich sollte das Verfahren nach dem „State of the Art" der Testkonstruktion (siehe z.B. Kemper/Ziegler/Krumm/Heene/Bühner 2015) entwickelt und psychometrisch überprüft werden. Die empirische Untersuchung gliederte sich demnach in drei Teile. In einer Vorstudie (siehe Abschnitt 2) wurden Indikatoren für „gute" und „schlechte" Präsentationsleistungen gesammelt, daraus Items formuliert und zu einem Testentwurf zusammengestellt. In Studie 1 (siehe Abschnitt 3) wurde auf der Grundlage von 193 Bewertungen studentischer Präsentationen die Faktorenstruktur der Testitems analysiert und wenig geeignete Items eliminiert. In Studie 2 (siehe Abschnitt 4) wurde die revidierte Fassung weiter optimiert und die Faktorstruktur überprüft, sowie weitere relevante Gütekriterien ermittelt (vgl. Kemper et al. 2015).

2. Itementwicklung (Vorstudie)

Der direkteste Weg, Indikatoren für gute und weniger gut gelungene Präsentationen zu finden, besteht darin, Dozenten und Studierende, die täglich in der Lehre mit Vorträgen konfrontiert sind, zu befragen (Marsh/Roche 1993; Rindermann 1999b). Als Grundlage für die Itemgenerierung wurden daher halbstandardisierte Interviews, angelehnt an die Critical-Incident-Technique nach Flanagan (1954), mit Studierenden unterschiedlicher Fachrichtungen der Gutenberg-Universität Mainz und der Rupprecht-Universität Heidelberg durchgeführt (N = 20 Studierende der Psychologie, Sprachwissenschaften, Geschichtswissenschaften, Anthropologie, Kunstgeschichte, und Politikwissenschaften zwischen 22 und 31 Jahren, M = 24 Jahre; 14 weiblich; im 2. bis 11. Fachsemester, M = 6.2). Die etwa halbstündigen Interviews verfolgten das Ziel, die aus studentischer Sicht erfolgskritischen Aspekte einer Präsentation zu erfassen. Dieser offene Zugang wurde explizit gewählt, um die Akzeptanz des Verfahrens bei der Zielgruppe zu erhöhen und eine Einengung der Indikatoren durch die Beschränkung auf ein bestimmtes theoretisches Modell zu vermeiden (vgl. Rindermann 1996). Auf der Grundlage der in den Interviews identifizierten Indikatoren wie z.B. *„Blickkontakt halten"* oder *„frei sprechen"*, wurden von den Autoren (TM, DO, CK) 74 Items entwickelt. Diese Items wurden anschließend durch sechs Items aus dem Heidelberger Inventar zur Lehrevaluation (HILVE; Rindermann/Amelang 1994) ergänzt. Als Antwortskala wurde eine vierstufige Ratingskala *(„trifft nicht zu" – „trifft eher*

nicht zu" – „trifft eher zu" – „trifft zu") gewählt. Am Ende des Fragebogens wurde ein fünfzehnstufiges Item eingefügt, das der Beurteilung der Gesamtleistung dient. Für dieses Item wurde die 15-Punkte-Notenskala der gymnasialen Oberstufe herangezogen. Der Testentwurf bestand aus insgesamt 81 Items.

3. Strukturanalyse des Testentwurfs und Itemreduktion (Studie 1)

3.1 Zielsetzung und Methoden

Ziel von Studie 1 war es, die Faktorstruktur des Testentwurfs zu ermitteln und ungeeignete Items zu eliminieren. Die Datenerhebung fand in zehn Seminaren des Psychologischen Instituts der Gutenberg-Universität Mainz statt. Die Seminare variierten sowohl über verschiedene Fachgebiete der Psychologie, als auch über den Studienzeitpunkt, an dem die Seminare typischerweise besucht werden (1. bis 9. Semester, *Md* = 3). In jedem der zehn Seminare wurde jeweils die Präsentation eines Referenten anhand des Evaluationsbogens bewertet. Eine Woche vor der Erhebung wurde ein Informationsblatt zu Zielen und Ablauf der Untersuchung ausgeteilt. Die Bewertung erfolgte nach Abschluss der Präsentation am Ende der Sitzung und wurde dabei sowohl von den Peers als auch vom leitenden Dozenten vorgenommen. In die Analyse gingen nur die 193 Präsentationsbewertungen der Studierenden (Peers) ein, nicht die Bewertungen der Dozenten.

3.2 Ergebnisse

Die Daten wurden einer explorativen Faktorenanalyse mit dem Statistikpaket SPSS 11.5 unterzogen, um Items mit inhaltlichen Ähnlichkeiten zu identifizieren. Es wurde eine Hauptachsenanalyse mit Promax-Rotation durchgeführt (vgl. Bühner 2011). Zur Ermittlung der Faktorenzahl wurden unterschiedliche in der Fachliteratur empfohlene Kriterien angelegt: Minimum Average Partial Test (MAP-Test; Velicer 1976), Parallelanalyse (Horn 1965) und Interpretierbarkeit der Lösungen (Bühner 2011; Zwick/Velicer 1986). Der MAP-Test legte acht Faktoren nahe. In der Parallelanalyse basierend auf 5000 Permutationen des Datensatzes (O'Connor 2000) lagen sechs Eigenwerte der Rohdaten über dem 95. Perzentil der Eigenwerte der zufälligen Permutationen. Da einige Studien nahelegen, dass die statistischen Kriterien tendenziell die Anzahl extrahierbarer Faktoren überschätzen (z.B. Zwick/Velicer 1986), wurden Lösungen mit fünf bis acht Faktoren extrahiert und die rotierten Ladungsmatrizen inspiziert. Fünf Faktoren erwiesen sich in allen berechneten Faktorlösungen als robust und interpretierbar. Zur Interpretation der Faktoren wurden deren Markiervariablen (Faktorladungen $\lambda > .60$) herangezogen. Auf dem ersten Faktor laden Items, die sich auf die Verständlichkeit und die korrekte Vermittlung des Stoffs beziehen, z.B. „Die Erklärungen des Referenten waren anschaulich" (Faktor *Korrekte und verständliche Vermittlung;* VER). Die Items zur Mediennutzung sind ebenfalls Teil des ersten Faktors. Dies ist inhaltlich plausibel, da die Verständlichkeit einer Präsentation auch davon abhängt, in welcher Art und Weise der Stoff präsentiert wird. Der zweite Faktor beinhaltet Aspekte, die sich auf die Begeisterung, Unterhaltung und Motivation der Hörerschaft beziehen, z.B. „Der Referent konnte Interesse für das Thema wecken" (Faktor *Motivieren und Interesse wecken;* MOT). Der dritte Faktor bezieht sich auf das Zeitmanagement der Präsentation, z.B. „Die Menge des vermittelten Stoffs war angemessen" (Faktor *Zeitmanagement;* ZEI). Dieser Faktor ist neben dem vierten Faktor inhaltlich am homogensten. Der vierte Faktor beinhaltet Items, die sich auf die Sprache, die Körperhaltung und die Gestik beziehen, z.B. „Der Referent hat frei vorgetragen" (Faktor *Vortragsstil;* STI). Der fünfte Faktor umfasst Aspekte der kritischen Auseinandersetzung mit dem Lehrstoff und der Einbeziehung des Publikums, z.B. „Der Referent hat Denkanstöße gegeben" (Faktor *Reflexion des Lehrstoffs;* REF).

Im Anschluss an die Ermittlung der Faktorstruktur der Testitems wurden wenig geeignete Items aus dem Testentwurf entfernt. Die Itemreduktion erfolgte nach statistischen und inhaltlichen Kriterien (vgl. Stanton/Sinar/Balzer/Smith 2002). Aufgrund ausreichend hoher Faktorladungen ($.27 < \lambda < .86$) der Items wurden inhaltliche Aspekte vorrangig beachtet: die Passung zum jeweiligen Faktor und die Passung zu den übrigen Items des Faktors. Für letztere wurde auf eine möglichst hohe Heterogenität der Items geachtet, um schmal operationalisierte Konstrukte zu vermeiden (vgl. Bühner 2011). Weiterhin wurden Items eliminiert, die substantielle Nebenladungen aufwiesen. Nach der Elimination psychometrisch ungeeigneter bzw. inhaltlich redundanter Items verblieben 33 Items in dieser Fragebogenversion.

4. Validierung des MPEP (Studie 2)

4.1 Zielsetzung und Methoden

Ziel der zweiten Studie war es, die in Studie 1 explorativ ermittelte fünffaktorielle Struktur zu bestätigen (faktorielle Validität) und weitere psychometrische Gütekriterien (Reliabilität, Kriteriumsvalidität) sowie Nutzen und Akzeptanz durch die Peers bzw. die Präsentierenden zu ermitteln. Zuvor wurde die Fragebogenversion aus Studie 1 weiter gekürzt. Dies wurde notwendig, da erste Erfahrungen im Lehrbetrieb vermuten ließen, dass diese Fassung aufgrund ihres Umfangs weniger häufig eingesetzt werden würde. Datengrundlage war eine zweite empirische Erhebung in sechs Seminaren des Psychologischen Instituts der Gutenberg-Universität, jeweils an zwei bzw. vier aufeinander folgenden Seminarterminen, in denen jeweils eine Präsentation gehalten und bewertet wurde.

Die Fragebogenversion aus Studie 1 enthielt 33 Ratingskalen-Items, ein Item für die Gesamtbewertung (15-Punkte-Notenskala der gymnasialen Oberstufe, fortan auch als Gesamtnote bezeichnet) und drei offene Fragen am Ende des Bogens[2], bei denen die Peers positive und negative Aspekte der Präsentation und Tipps für den/die Präsentierenden frei formulieren konnten. Im Gegensatz zur ersten Erhebung gab es zwei Versionen des Fragebogens, die sich in den Instruktionen und in den Itemformu-

[2] Während der Interviews bezeichneten Studierende Items mit offenem Antwortformat als besonders informativ und nützlich. Daher wurden diese in die revidierte und finale Fassung des MPEP aufgenommen.

lierungen unterschieden: eine „Einzelversion" und eine „Gruppenversion". Während die Einzelversion zur Bewertung einer einzelnen Leistung, also der Leistung nur einer Person der präsentierenden Studierendengruppe, herangezogen wurde, diente die Gruppenversion der Bewertung der gesamten Studierendengruppe. Zwei Gründe waren für die Entwicklung und Überprüfung einer Gruppenversion ausschlaggebend. Zum Einen ist diese bei der Bewertung von Gruppen ökonomischer im Vergleich zur Einzelversion, die für jede Person ausgefüllt werden muss. Zum Anderen entspricht die Gruppenversion dem in der Praxis oft beobachteten diagnostischen Prozess der Veranstaltungsleiter. Wenn eine Präsentationsleistung zu Teilen in die Gesamtnote eingeht, wird aus pragmatischen Gründen oft die Gruppenleistung bewertet anstatt alle Einzelleistungen zu bewerten. Es ist unmittelbar evident, dass diese Mittelwertbildung, insbesondere bei hoher Leistungsvarianz, dem Einzelnen nicht gerecht wird. Dennoch sollte überprüft werden, ob eine solche ökonomische Rückmeldung der Präsentationsleistung ein Mindestmaß an verwertbaren Informationen für die Präsentierenden mit sich bringt, um ihren Einsatz zu rechtfertigen. Damit die Präsentierenden sich im Hinblick auf die zurückgemeldete Gruppenleistung einordnen können, wurde der Gruppenversion ein zusätzliches Item beigefügt, auf dem die „Unterschiedlichkeit" der Einzelleistungen bewertet werden kann.

Eine Woche vor der Erhebung wurde in jedem Seminar ein Informationsblatt zum Ablauf und typischen Beurteilerfehlern ausgeteilt. Die Präsentierenden einer Sitzung wurden anhand der Einzel- oder anhand der Gruppenversion bewertet. Dabei wurde die Version innerhalb einer Lehrveranstaltung von Präsentation zu Präsentation alterniert. Die Bewertungen wurden durch die Peers jeweils am Ende der Lehrveranstaltung vorgenommen. Insgesamt wurden 501 Bewertungen vorgenommen, die sich auf zwei voneinander unabhängige Stichproben verteilen (Einzelversion N = 325, Gruppenversion N = 176). Zusätzlich haben auch die Veranstaltungsleiter die Präsentationen bewertet.

Neben den Bewertungen der Präsentationsleistungen wurden Nachbefragungen mit den Peers (N = 107) und den Präsentierenden (N = 32) durchgeführt, um Nutzen und Akzeptanz der Peer-Evaluation zu ermitteln. Hierzu wurde ein kurzer Fragebogen zusammengestellt, der nach dem Ende der Erhebung von den Peers bzw. den Präsentierenden ausgefüllt wurde. Die Präsentierenden bekamen zuvor eine detaillierte Rückmeldung ihrer Leistung (in Form eines PDF-Dokuments) auf der Grundlage der von den Peers vorgenommenen Bewertungen. Die Rückmeldung enthielt eine Erklärung zum Aufbau und den verwendeten Graphiken und statistischen Kennwerten. Für die verhaltensnahen Items, die fünf aggregierten Skalenwerte und für die Gesamtnote wurden die Häufigkeitsverteilung der Antworten, die Prozentränge der Antwortkategorien, Mittelwert und Streubreite in graphischer Form und die Anzahl der Evaluationen, die fehlenden Werte, Mittelwert und Median in Textform rückgemeldet. Zudem waren die freien Antworten der Peers zu positiven und negativen Aspekten angehängt. Im Folgenden werden die Ergebnisse zu den psychometrischen Gütekriterien des MPEP näher beschrieben.

4.2 Ergebnisse

4.2.1 Faktorielle Validität

Aufgrund erster Erfahrungen in der Lehre erwies sich die Fassung mit 33 Ratingskalen-Items aus Studie 1 als zu lang. Daher wurde der Testentwurf (Einzelversion des Fragebogens) nach den bereits in Studie 1 verwendeten Kriterien auf vier Items pro Skala, also insgesamt 20 Items, reduziert. Anhand der Daten einer unabhängigen Stichprobe (Gruppenversion) wurde dann die faktorielle

Tab. 1: Faktorladungen (λ) und deskriptive Statistiken der Items für die Gruppenversion und die Einzelversion (Studie 2).

Item		Gruppenversion			Einzelversion		
		λ	M	SD	λ	M	SD
	Faktor 1: Korrekte und verständliche Vermittlung						
2	Die Referenten haben den Stoff des Referats selbst verstanden	.51	3.54	.54	.66	3.47	.73
4	Die Erklärungen des Referenten waren anschaulich	.67	3.24	.66	.77	3.20	.80
13	Die Referenten vermittelten, in welchem Zusammenhang die Informationen zu sehen sind	.65	3.18	.73	.63	3.14	.73
20	Die Folien waren eine sinnvolle Ergänzung des Vortrags	.69	3.25	.81	.62	3.29	.73
	Faktor 2: Motivieren und Interesse wecken						
9	Der Vortrag wirkte abwechslungsreich	.84	2.90	.86	.79	2.66	.79
11	Ich wurde zum Mitdenken motiviert	.67	2.86	.86	.74	2.95	.74
12	Die Referenten konnten Interesse für das Thema wecken	.78	3.10	.80	.75	3.02	.82
16	Das Thema wurde spannend präsentiert	.88	2.87	.93	.82	2.65	.82
	Faktor 3: Zeitmanagement						
7	Die Referenten haben zu viele Details präsentiert	.62	3.17	.87	.58	3.31	.77
14	Die Menge des vermittelten Stoffs war angemessen	.76	3.33	.79	.74	3.38	.66
18	Das Referat hatte die richtige Länge	.64	3.35	.78	.70	3.31	.75
19	Die Referenten haben sich beim Inhalt auf das Wichtigste konzentriert	.73	3.16	.71	.69	3.24	.62
	Faktor 4: Vortragsstil						
1	Die Referenten haben sich um eine lockere, angenehme Atmosphäre bemüht	.80	3.42	.75	.77	3.07	.82
10	Die Referenten hatten meistens Blickkontakt mit den Zuhörern	.56	3.11	.79	.65	3.01	.89
6	Die Gestik der Referenten hat den Vortrag unterstrichen	.78	2.86	.90	.78	2.64	.85
15	Die Referenten haben frei vorgetragen	.68	3.16	.91	.71	2.96	.88
	Faktor 5: Reflexion des Lehrstoffs						
3	Kritische Punkte der behandelten Themen wurden hervorgehoben	.58	2.96	.79	.76	2.96	.81
5	Die Referenten haben Diskussionen angeregt	.54	2.11	.95	.67	2.68	1.01
8	Die Referenten haben Zeit für Fragen gelassen	.38	3.34	.75	.53	3.29	.79
17	Die Referenten haben Denkanstöße gegeben	.79	2.71	.84	.75	2.76	.81

Validität der 20-Item-Version überprüft, indem eine konfirmatorische Faktorenanalyse (MLR-Schätzung) mit MPlus Version 5.2 durchgeführt wurde. Ein fünffaktorielles Modell mit korrelierten Faktoren wurde spezifiziert. In unserer Studie bewerteten immer mehrere Studierende dieselbe Präsentation. Es handelt sich also um ein Design mit zwei Ebenen, wobei die Beurteiler Ebene 1 repräsentieren und die Präsentationen Ebene 2. Um diese Abhängigkeitsstruktur in den Daten zu kontrollieren, wurde die Option TYPE = COMPLEX mit der Präsentation als Clustervariable verwendet (Huang/Cornell 2016). Die Beurteilung der Modellpassung beruht auf einer Auswahl von Fitindizes (Beauducel/Wittmann 2005), die anhand der Kriterien von Hu/Bentler (1999) bewertet wurden. Das in Studie 1 explorativ ermittelte und in Studie 2 anhand der Daten der Einzelversion optimierte fünffaktorielle Testmodell des MPEP zeigt eine akzeptable Passung auf die mit der Gruppenversion erhobenen Daten ($\chi^2(160) = 268.7$, $p < .001$, RMSEA = .061, SRMR = .058, CFI = .939). Alle Items weisen substantielle Ladungen auf den Faktoren auf, denen sie zugeordnet waren. Die Kreuzvalidierung des fünffaktoriellen Testmodells in einer unabhängigen Stichprobe ist demnach gelungen. Die Faktorladungen der Items (Gruppenversion) sind in Tabelle 1 aufgeführt (zu Vergleichszwecken werden dort ebenfalls die Faktorladungen für die Einzelversion angegeben, $\chi^2(160) = 323.1$, $p < .001$, RMSEA = .056, SRMR = .050, CFI = .953).
Die Interkorrelationen der fünf Faktoren liegen zwischen .56 *(Zeitmanagement mit Vortragsstil)* und .92 *(Korrekte und verständliche Vermittlung* mit *Motivieren und Interesse wecken)*. Die Faktoren des MPEP weisen demnach eine mittlere bis sehr hohe Ähnlichkeit zueinander auf, die insbesondere die beiden Faktoren *Korrekte und verständliche Vermittlung* und *Motivieren und Interesse wecken* betreffen. Allerdings weisen diese auch zu 15% ungeteilte Varianzanteile auf, die gegen eine Identität der beiden Evaluationsaspekte sprechen.

4.2.2 Reliabilität

Als Schätzer für die Reliabilität wurde das gängige Cronbach α und McDonalds Ω_w (McDonald 1999) verwendet. Es wurden keine wesentlichen Unterschiede zwischen Einzel- und Gruppenversionen beobachtet (siehe Tabelle 2). Für alle Skalen des MPEP, mit Ausnahme der Skala *Reflexion des Lehrstoffs* der Gruppenversion, lagen die Reliabilitätskoeffizienten in einem zufriedenstellenden bis guten Bereich.

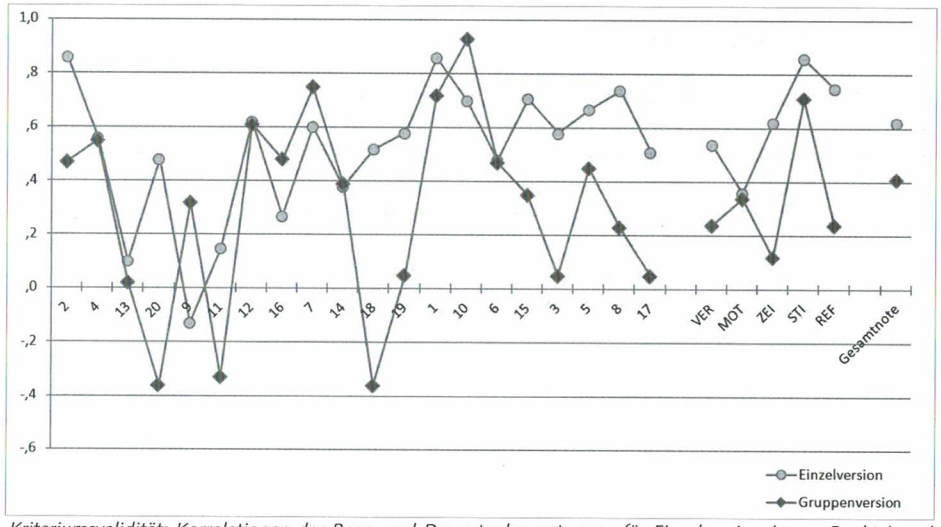
4.2.3 Kriteriumsvalidität

Die Kriteriumsvalidität des MPEP wurde anhand der Übereinstimmung zwischen den Bewertungen der Dozenten auf dem MPEP und den pro Präsentation aggregierten Bewertungen der Peers korrelativ bestimmt (Einzelversion: N = 17, Gruppenversion: N = 9). Korrelationskoeffizienten wurden gewählt, um die Befunde mit Werten aus einer Metaanalyse von Falchikov/Goldfinch (2000) vergleichen zu können. Dort lagen die Korrelationskoeffizienten zwischen den Peer- und Dozentenbewertungen zwischen r = .14 und .99 (M = .69).
In unserer Studie lagen die Koeffizienten auf Skalenebene in einem Bereich von r = .36 bis .86 (siehe Abbildung 1) und somit in dem von der Metaanalyse angegebenen Rahmen. Die Skala *Vortragsstil* enthält sehr verhaltensnahe und gut beobachtbare Indikatoren wie Blickkontakt, freies Vortragen, Gestik etc., und wies mit r = .86 die höchste Korrelation auf. Für die Gesamtnote zeigte sich eine Korrelation von r = .62. Auf Itemebene lagen 80% der Korrelationskoeffizienten für die Einzelversion über r = .40, was üblicherweise als befriedigende Validität angesehen wird (Bühner 2011). Der Median der Koeffizienten betrug r = .57.
Für die Gruppenversion lagen die Korrelationskoeffizienten auf Skalenebene zwischen r = .12 (Skala *Zeitmanagement)* und .71 (Skala *Vortragsstil)*. Für die Gesamtnote betrug die Korrelation zwischen Dozenten und Peers r = .41. Wie aus Abbildung 1 ersichtlich, lagen auf Itemebene weniger Koeffizienten in dem aufgrund der Metaanalyse erwarteten Bereich als für die Einzelversion. Der Median lag nur noch bei r = .37. Etwa ein Drittel der Koeffizienten lag außerhalb des von der Metaanalyse angegebenen Wertebereichs. Drei Items wiesen sogar negative Koeffizienten auf, die sich aufgrund schiefer Verteilung

Tab. 2: Reliabilitätsschätzungen (McDonald Ω_w/Cronbach α) für die Skalenwerte der Einzel- und Gruppenversion des MPEP.

Skala	Einzelversion	Gruppenversion
Korrekte und verständliche Vermittlung	.78/.77	.76/.74
Motivieren und Interesse wecken	.86/.86	.89/.88
Zeitmanagement	.78/.76	.80/.78
Vortragsstil	.83/.82	.83/.80
Reflexion des Lehrstoffs	.80/.77	.72/.65

Abb. 1: Kriteriumsvalidität



Kriteriumsvalidität: Korrelationen der Peer- und Dozentenbewertungen für Einzelversion (graue Punkte) und Gruppenversion (schwarze Quadrate), auf Ebene von Items (1-20, sortiert nach Skala), Skala (VER = Korrekte und verständliche Vermittlung, MOT = Motivieren und Interesse wecken, ZEI = Zeitmanagement, STI = Vortragsstil, REF = Reflexion des Lehrstoffs) und Gesamtnote.

lungen, stark unterschiedlicher Varianzen und Ausreißer bei geringem Stichprobenumfang ergaben. Insgesamt 50% der Korrelationskoeffizienten lagen noch über .40 (siehe Abbildung 1).

4.2.4 Nutzen und Akzeptanz

In der Nachbefragung wurden verschiedene Nutzen- und Akzeptanzaspekte zum MPEP mit Ratingskalen-Items erfasst; bei den Peers z.B. *„Ich fände es gut, über meine Leistung bei einem eigenen Referat ein Feedback von meinen Kommilitonen zu erhalten"* oder *„Das Bewerten der Referenten hat mich zum Nachdenken über meine eigene Referats-Leistung angeregt"*. Die Antworten erfolgten auf derselben vierstufigen Ratingskala von *„trifft nicht zu"* (1) bis *„trifft zu"* (4) wie im MPEP. Alle Itemmittelwerte lagen signifikant über der neutralen Skalenmitte von 2.5 (*t*-Test für eine Stichprobe, alle *p*-Werte < .01, zweiseitig). Die Peer-Evaluation wurde demnach von den Bewertern insgesamt positiv beurteilt. Die Handhabung des Fragebogens (Item: *„Es fiel mir leicht, die Leistung meiner Kommilitonen mit der Einzelversion bzw. Gruppenversion einzuschätzen")* wurde für die Einzelversion mit $M = 3.12$ (SD = 0.69) positiver bewertet als für die Gruppenversion ($M = 2.72$, SD = 0.84), $t(103) = 4.60$, $p < .001$, Cohens (1988) $d_z = 0.45$. Die Angemessenheit des Zeitaufwands der Einzelversion ($M = 2.84$, SD = 0.85) wurde allerdings signifikant niedriger eingestuft als die Angemessenheit des Zeitaufwands der Gruppenversion ($M = 3.18$, SD = 0.73), $t(103) = 4.41$, $p < .001$, $d_z = 0.43$. Bei einem Item aus der Nachbefragung sollten die Studierenden angeben, welche Version des Peer-Feedbacks sie für ihre eigene Präsentation bevorzugen würden. Ungefähr 80% entschieden sich dabei für die Einzelversion und gegen die Gruppenversion, $\chi^2(1) = 37.1$, $p < .001$.

In der Nachbefragung der Präsentierenden wurden ebenfalls verschiedene Nutzen- und Akzeptanzaspekte mit Ratingskalen-Items erfasst, z.B. *„Ich würde das Peer-Feedback auch gerne bei meinen zukünftigen Referaten erhalten"* oder *„Das Peer-Feedback hat mir meine Stärken und Schwächen aufgezeigt"*. Die Itemmittelwerte lagen sowohl bei den mit der Einzelversion als auch bei den mit der Gruppenversion bewerteten Präsentierenden bis auf wenige Ausnahmen signifikant über der neutralen Mitte der Antwortskala von 2.5 ($p < .05$, zweiseitig). Dies zeigt eine insgesamt eher positive Bewertung des Peer-Feedbacks. Allerdings lagen die Mittelwerte der Einschätzungen von Nutzen und Akzeptanz bei der Gruppenversion stets unter denen der Einzelversion. Um zu untersuchen, ob die Version des MPEP einen substantiellen Einfluss auf die Einschätzung von Nutzen und Akzeptanz des Peer-Feedbacks seitens der Präsentierenden hat, wurde eine multivariate Varianzanalyse über die Items mit dem Between-subjects-Faktor Version (Einzelversion, Gruppenversion) berechnet. Es gab keinen signifikanten Effekt des Faktors Version, $F(14, 17) = 1.43$, $p = .24$.

5. Diskussion

Ziel der Untersuchung war es, ein neues, inhaltlich verankertes, ökonomisches und psychometrisch validiertes Instrument zur Peer-Evaluation studentischer Präsentationen zu entwickeln, das für den standardmäßigen Einsatz in universitären Lehrveranstaltungen geeignet ist und Studierende durch eine detaillierte und diagnostisch gehaltvolle Rückmeldung bei der Verbesserung ihrer Präsentationskompetenz unterstützt. In einer Serie von drei empirischen Erhebungen wurden Indikatoren für „gute" und „schlechte" Präsentationen identifiziert und zu einem Fragebogen zusammengestellt, der anschließend optimiert und psychometrisch überprüft wurde. Die beiden Versionen des MPEP, die Einzel- und die Gruppenversion, erwiesen sich als hinreichend reliabel, inhaltsvalide und faktoriell valide. Die Einzelversion zeigte im Vergleich der Bewertungen der Peers und der Dozenten zudem eine akzeptable bis hohe Kriteriumsvalidität. Sie wird von den potentiellen Anwendern, den Peers und den Präsentierenden, angenommen und als nützlich eingeschätzt. Im Hinblick auf die letztgenannten Aspekte der Güte schneidet die Gruppenversion, die der Einzelversion in der Ökonomie klar überlegen ist, allerdings etwas schlechter ab.

Das MPEP ermöglicht es Studierenden, eine über die Bewertung der Veranstaltungsleiter hinausgehende Rückmeldung ihrer Präsentationsleistung zu bekommen und schafft somit eine weitere Grundlage für die Verbesserung ihrer Präsentationskompetenz (vgl. Frese/Zapf 1994; Gibbs/Simpson 2004; Nicol/Macfarlane-Dick 2006). Zudem scheint das Peer-Feedback nicht nur eine zum Dozenten-Feedback inkrementelle Informationsquelle zu sein, sondern auch eher verhaltensmodifizierend zu wirken als das Feedback von Dozenten oder anderem Lehrpersonal (DeNisi/Randolph/Blencoe 1982). Die Studierenden erhalten durch das MPEP eine detaillierte Rückmeldung ihrer Präsentationsleistung auf der Ebene von Items und Skalenwerten, eine Gesamtbewertung und frei formulierte Hinweise zu positiven und negativen Aspekten der erbrachten Leistung. Sie bekommen Hinweise zu Schwächen in bestimmten Bereichen der Präsentationskompetenz, z.B. ob ihre fachlichen und didaktischen Kompetenzen ausreichen, um den Lehrstoff verständlich zu vermitteln, oder ob sie die Adressaten in Zukunft stärker einbeziehen und eine kritische Auseinandersetzung mit dem Lehrstoff anregen müssen. Im Gegensatz zu früheren Verfahren (z.B. Falchikov 1995; Langan et al. 2008; MacAlpine 1999) werden allerdings nicht nur Rückmeldungen für „abstrakte" Kompetenzbereiche wie *Korrekte und verständliche Vermittlung* oder *Reflexion des Lehrstoffs* gegeben. Das MPEP gibt auch Rückmeldungen auf Ebene der Einzelitems, die sich meist auf konkrete und erfolgskritische Verhaltensweisen beziehen (z.B. *„Der Referent hat Denkanstöße gegeben")* und somit gute Ausgangspunkte für eine Verbesserung der Kompetenz darstellen (Gibbs/Simpson 2004). Das Item Gesamtnote bildet die Leistung des Präsentierenden in einem Indikator ab, so wie sie von den Adressaten der Präsentation zusammenfassend bewertet wurde. Die Antwortskala orientiert sich am Notensystem der gymnasialen Oberstufe. Den Feedback-Empfängern ist dieses Bewertungssystem somit vertraut und erleichtert ihnen ihre Gesamtleistung einzuordnen. Die drei offenen Fragen am Ende des Evaluationsbogens erlauben die Rückmeldung von Aspek-

ten, die sich nicht in das „Korsett" der Ratingskalen pressen lassen und von den Präsentierenden üblicherweise als besonders informativ eingeschätzt werden (für Details siehe Mitschke 2009). Das MPEP ermöglicht eine differenzierte und umfassende Rückmeldung von Präsentationsleistungen.

Zudem wurde das MPEP nach dem „State of the Art" der Testkonstruktion entwickelt und umfassend psychometrisch überprüft. Durch die Einbindung derjenigen (Studierende) in den Konstruktionsprozess, die „gute" und „schlechte" Präsentationen täglich erleben, wurde die Inhaltsvalidität des Evaluationsverfahrens sichergestellt. Die faktorielle Validität des fünffaktoriellen Testmodells konnte empirisch gesichert werden. Die Struktur der entwickelten Testitems wurde zunächst explorativ ermittelt. Anschließend wurden nach allseits empfohlenen Methoden redundante und ungeeignete Items eliminiert und eine ökonomische, 20 Items umfassende finale Version erstellt, deren Struktur über eine konfirmatorische Faktoranalyse belegt werden konnte. Die Reliabilität der faktorenanalytisch ermittelten Skalen ist bei beiden Versionen für den angestrebten Verwendungszweck ausreichend. Auch die Kriteriumsvalidität des MPEP konnte gesichert werden. Die Bewertungen der Peers und der Dozenten stimmten, insbesondere für die Einzelversion, in einem Maße überein, das aufgrund metaanalytischer Befunde (Falchikov/Goldfinch 2000) erwartet werden kann. Höhere Übereinstimmungskoeffizienten als die beobachteten, sind durch die unterschiedlichen Perspektiven von Peers und Dozenten und damit einhergehende Unterschiede in Beobachtungs- und Beurteilungsfehler, nicht zu erwarten. Im Vergleich zur Einzelversion waren die Korrelationskoeffizienten der Gruppenversion, mit einer Ausnahme (Item 9), insgesamt niedriger. Anhand dieser Differenzen wird deutlich, dass die Aggregation von mehreren Bewertungen negative Konsequenzen im Hinblick auf die Güte mit sich bringt. Dieser Befund ist vor dem Hintergrund des mit der Aggregation einhergehenden Informationsverlusts notwendigerweise trivial. Entscheidend ist in diesem Zusammenhang die Frage, ob der Verlust an psychometrischer Güte auf der einen Seite durch den Zugewinn an Ökonomie und somit Akzeptanz der Veranstaltungsleitung auf der anderen Seite aufgewogen werden kann.

Die Nachbefragung zeigte, dass das Peer-Feedback sowohl von denjenigen, die es geben, als auch von denjenigen, die es empfangen, als sinnvoll und nützlich erachtet wird. Sämtliche Akzeptanz- und Nutzenaspekte wurden positiv eingeschätzt. Die Einschätzungen unterschieden sich allerdings im Hinblick auf die Art des Feedbacks. Die Peers schätzten die Handhabbarkeit der Einzelversion im Vergleich zur Gruppenversion als höher ein. Dies liegt nahe, da bei der Gruppenversion mehrere Bewertungen zu einem Urteil zusammengefasst werden müssen. Dieser Prozess ist diagnostisch ungleich aufwendiger als eine einzelne Person zu bewerten. Im Gegensatz zur Einzelversion wird die Gruppenversion vorteilhafter im Bezug auf die Ökonomie angesehen. Mit der Gruppenversion muss nur eine Bewertung abgegeben werden, anstelle von mehreren Bewertungen wie im Fall der Einzelversion. Trotz positiverer Einschätzung der Gruppenversion bei der Ökonomie äußerten die Studierenden eine klare Präferenz für die Einzelversion. Vier Fünftel gaben bspw. an, für die eigene Präsentation Peer-Feedback auf Grundlage der Einzelversion des MPEP erhalten zu wollen, wenn sie wählen könnten. Angesichts des höheren diagnostischen Gehalts dieser Art von Rückmeldung ist diese Wahl verständlich.

Auf Grundlage der vorgelegten Befunde kann der Einsatz der Einzelversion des MPEP für die Peer-Evaluation studentischer Präsentationsleistungen in der universitären Lehre empfohlen werden. Die Rückmeldung ist hinreichend informativ, um Studierende bei der Verbesserung ihrer Präsentationskompetenz zu unterstützen. Außerdem hat sich das Verfahren psychometrisch bewährt und erlaubt eine reliable und valide Rückmeldung von guten bzw. weniger guten Präsentationsleistungen. Allerdings erfordert der Einsatz des Evaluationsverfahrens auch eine „Investition", insbesondere von der Veranstaltungsleitung. Diese muss zunächst entscheiden, auf wie viel Lehrveranstaltungszeit verzichten werden kann, um die Evaluationen durchzuführen. Eine Verbesserung der Qualität der Präsentationen und positive Effekte auf die Veranstaltungsevaluation sind attraktive Anreize für die Veranstaltungsleitung. Der Zeitaufwand für die Evaluation mittels MPEP liegt bei ca. drei bis vier Minuten. In Veranstaltungen mit einer höheren Anzahl von Präsentierenden pro Veranstaltungstermin könnte, als zeitlich ökonomischere Alternative, die Gruppenversion eingesetzt werden. Deren Einsatz kann allerdings aufgrund des geringeren diagnostischen Gehalts der Rückmeldung und der geringeren Kriteriumsvalidität (im Sinne der Übereinstimmung von Peers und Dozenten) nur mit Einschränkungen empfohlen werden. Da der Nutzen der Gruppenversion jedoch sowohl von den Peers als auch von den Präsentierenden positiv bewertet wurde, wenn auch in geringerem Maße als die Einzelversion, ist ein Einsatz der Gruppenversion sicherlich vorteilhafter als der vollständige Verzicht einer Evaluation. In diesem Fall bietet es sich allerdings an, wegen der geringeren Kriteriumsvalidität, eine ergänzende Rückmeldung der Präsentationsleistung durch den Dozenten vorzunehmen.

Eine mögliche methodische Einschränkung unserer Untersuchung liegt im Design unserer Studien, in der immer mehrere Studierende dieselbe Präsentation bewerteten, entsprechend dem Ablauf von Lehrveranstaltungen. Wir untersuchten also relativ große Stichproben auf Ebene 1, das Design beinhaltete jedoch relativ wenige Beobachtungseinheiten (Präsentationen) auf Ebene 2. Für die explorativen Faktorenanalysen in Studie 1 könnten also die aufgrund der Schachtelung von Beurteilern innerhalb von Präsentationen hervorgerufenen Abhängigkeiten zu Überschätzungen der Inter-Item-Korrelationen geführt haben. Für die konfirmatorische Faktorenanalyse in Studie 2 stand uns ein statistisches Verfahren zur Verfügung, das als verhältnismäßig robust gegenüber Abhängigkeit im Stichprobendesign angesehen wird (TYPE = COMPLEX Option in MPlus; Asparouhov 2005).

Das entwickelte MPEP ermöglicht es Studierenden durch das Üben und Reflektieren des selbständigen Lernens, der Vermittlung von Inhalten, des sicheren Umgangs mit audiovisuellen Medien und des freien Sprechens ihre Präsentationskompetenz zu verbessern. Da

die mediengestützte Präsentation heute der zentrale Modus zur Vermittlung von Arbeitsergebnissen ist (vgl. Schnack 2011), bereitet die Peer-Evaluation Studierende auf zentrale Anforderungen der Arbeitswelt vor, die nach der Graduierung auf sie warten.

Literaturverzeichnis

Asparouhov, T. (2005): Sampling weights in latent variable modeling. In: Structural Equation Modeling – a Multidisciplinary Journal, Vol. 12/No. 3, pp. 411-434.

Beauducel, A./Wittmann, W.W. (2005): Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. In: Structural Equation Modeling, Vol. 12/No. 1, pp. 41-75.

Bromme, R./Rambow, R. (1993): Verbesserung der mündlichen Präsentation von Referaten: Ein Ausbildungsziel und zugleich ein Beitrag zur Qualität der Lehre. In: Das Hochschulwesen, Jg. 41/H. 6, S. 289-297.

Bühner, M. (2011): Einführung in die Test- und Fragebogenkonstruktion. München.

Cohen, J. (1988): Statistical power analysis for the behavioral sciences (2nd ed.). Mahwah: Erlbaum Associates.

DeNisi, A.S./Randolph, W.A./Blencoe, A.G. (1982): Level and source of feedback as determinants of feedback effectiveness. In: Academy of Management Journal, Vol. 3, pp. 175-179.

Falchikov, N. (1995): Peer Feedback marking: developing peer assessment. In: Innovations in Education and Training international, Vol. 23, pp. 175-187.

Falchikov, N./Goldfinch, J. (2000): Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. In: Review of Educational Research, Vol. 70/No. 3, pp. 287-322.

Flanagan, J.C. (1954): The Critical Incident Technique. In: Psychological Bulletin 51, 1954, 4, pp. 327-358.

Frese, M./Zapf, D. (1994): Action as the core of work psychology: A German approach. In: Triandis, H.C./Dunnette, M.D./Hough, L. (eds.): Handbook of industrial and organizational psychology, Vol. 4. Palo Alto, California: Consulting Psychologists Press, pp. 271-340.

Gibbs, G./Simpson, C. (2004): Conditions under which assessment supports students' learning In: Learning and Teaching in Higher Education, Vol. 1, pp. 3-31.

Gold, A./Souvignier, E. (2000): Rückmeldegespraeche nach studentischen Referaten: Ein Beitrag zur Verbesserung von Lehre? Feedback for students' oral presentations. A contribution to the improvement of education? In: Krampen/Zayer (Hrsg.): Psychologiedidaktik und Evaluation II. Neue Medien, Psychologiedidaktik und Evaluation in der psychologischen Haupt- und Nebenfachausbildung. Bonn, S. 203-218.

Horn, J.L. (1965): A rationale and test for the number of factors in factor analysis. In: Psychometrika, Vol. 30/No. 2, pp. 179-185.

Hu, L.T./Bentler, P.M. (1999): Cutoff criteria of fit indexes in covariance structure analysis: Convencional criteria versus new alternatives. In: Structual Equation Modeling, Vol. 6/No. 1, pp. 1-55.

Huang, F.L., & Cornell, D.G. (2016): Using multilevel factor analysis with clustered data investigating the factor structure of the Positive Values Scale. In: Journal of Psychoeducational Assessment, Vol. 43/No. 1, pp. 3-14.

Kemper, C.J./Ziegler, M./Krumm, S./Heene, M./Bühner, M. (2015): Testkonstruktion In: Stemmler, Margraf-Stiksrud (Hg.): Lehrbuch Diagnostik. Bern, S. 157-221.

Langan, A./Shukur, D.M./Cullen, W./Penney, D./Preziosi, R.F./Wheater, C. (2008): Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. In: Assessment and Evaluation in Higher Education, Vol. 33/No. 2, pp. 179-190.

MacAlpine, J.M.K. (1999): Improving and Encouraging Peer Assessment of Student Presentations. In: Assessment and Evaluation in Higher Education, Vol. 24/No. 1, pp. 15-25.

McDonald, R.P. (1999): Test theory: A unified treatment. Erlbaum.

Marsh, H.W.; Roche, L.A. (1993): The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. In: American Educational Research Journal, Vol. 30, pp. 217-251.

Mitschke, T. (2009): Entwicklung eines Fragebogens für die Peer-Evaluation von studentischen Referaten. Master thesis, Johannes Gutenberg-Universität Mainz.

Nicol, D.J./Macfarlane Dick, D. (2006): Formative assessment and self regulated learning: A model and seven principles of good feedback practice. In: Studies in higher education 31, 2006, 2, S. 199-218

O'Connor, B.P. (2000): SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. In: Behavior Research Methods Instruments & Computers, Vol. 32/No. 3, pp. 396-402.

Preiser, S. (1995): Feedback nach Referaten. Ein Beitrag zur Verbesserung der Lehre. In: Das Hochschulwesen, Jg. 43/H. 2, S. 114-116.

Rindermann, H. (1996): Zur Qualität studentischer Lehrveranstaltungsevaluationen: Eine Antwort auf Kritik an der Lehrevaluation. In: Zeitschrift für Pädagogische Psychologie, Jg. 10, S. 129-145.

Rindermann, H. (1999a): Bedingungs- und Effektvariablen in der Lehrevaluationsforschung: Konzeption und Prüfung des Münch multifaktoriellen Modells der Lehrveranstaltungsqualität. In: Unterrichtswissenschaft, Jg. 27/H. 4, S. 357-380.

Rindermann, H. (1999b): Was zeichnet gute Lehre aus? Ergebnisse einer offenen Befragung von Studierenden und Lehrenden In: Zeitschrift für Hochschuldidaktik, Jg. 23/H. 1, S. 136-156.

Rindermann, H. (2003): Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. In: Zeitschrift für Evaluation, H. 2, S. 233-256.

Rindermann, H./Amelang, M. (1994): Entwicklung und Erprobung eines Fragebogens zur studentischen Veranstaltungsevaluation. In: Empirische Pädagogik, Jg. 8/H. 2, S. 131-151.

Saito, H./Fujita, T. (2004): Characteristics and user acceptance of peer rating in EFL writing classrooms. In: Language Teaching Research, Vol. 8/No. 1, pp. 31-54.

Schnack, J. (2011): Die Präsentation als komplexe Leistung. Oder: Was man lernt, wenn man Präsentieren lernt. In: Pädagogik, Jg. 63/H. 12, S. 6-9.

Stanton, J.M./Sinar, E.F./Balzer, W.K./Smith, P.K. (2002): Issues and strategies for reducing the length of self-report scales. In: Personnel Psychology, Vol. 55, pp. 167-194.

Topping, K. (1998): Peer assessment between students in colleges and universities. In: Review of Educational Research, Vol. 68/No. 3, pp. 249-276.

Topping, K. (2009): Peer Assessment. In: Theory into Practice, Vol. 48/No. 1, pp. 20-27.

Topping, K.J. (2010): Methodological quandaries in studying process and outcomes in peer assessment. In: Learning and Instruction, Vol. 20/No. 4, pp. 339-343.

Velicer, W.F. (1976): Determining the number of components from a matrix of partial correlations. In: Psychometrika, Vo. 41/No. 3, pp. 321-327.

Wippermann, B. (2014): Die Präsentationsleistung und die Präsentationsprüfung in der Profiloberstufe – Handreichung und Arbeitshilfe für die Schulen. In: Behörde für Arbeit, Soziales, Familie und Integration Hamburg (Hg.): LI-Handreichung. Hamburg.

Zwick, W.R./Velicer, W.F. (1986): Comparison of Five Rules for Determining the Number of Components to Retain. In: Psychological Bulletin, Vol. 99/No. 3, pp. 432-442.

■ **Dr. Christoph J. Kemper,** Senior Researcher, Insitute of Cognitive Science and Assessment & Luxembourg Centre for Educational Testing (LUCET), University of Luxembourg,
E-Mail: christoph.kemper@uni.lu

■ **Tina Mitschke,** Dipl.-Psych., Studienbüro Bildungswissenschaften, Johannes Gutenberg-Universität Mainz,
E-Mail: tina.mitschke@uni-mainz.de

■ **Dr. Wolfram Rollett,** Professur für „Empirische Bildungsforschung mit dem Schwerpunkt Schulentwicklung" Pädagogische Hochschule Freiburg, Institut für Erziehungswissenschaft,
E-Mail: wolfram.rollett@ph-freiburg.de

■ **Viktoria Kemper,** Dipl. Soz.-Wiss., Wiss. Mitarbeiterin, Fachgebiet Information Systems and Electronic Services, Technische Universität Darmstadt,
E-Mail: kemper@ise.tu-darmstadt.de

■ **PD Dr. Daniel Oberfeld,** Senior Researcher Psychologisches Institut, Johannes Gutenberg-Universität Mainz,
E-Mail: oberfeld@uni-mainz.de