Empirische Sonderpädagogik, 2016, Nr. 2, S. 119-139
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Wie lässt sich Klassenführungsexpertise messen? Überprüfung eines videobasierten Erhebungsinstruments für Lehrkräfte unter Anwendung der Generalisierbarkeitstheorie

Gino Casale, Sarah Strauß, Thomas Hennemann & Johannes König

Universität zu Köln

Zusammenfassung

Klassenführungsexpertise stellt ein wesentliches Merkmal der professionellen Kompetenz von Lehrkräften dar und geht mit positiven Wirkungen sowohl hinsichtlich der Qualität des Unterrichts als auch im Kontext sonderpädagogischer Förderung einher. Um kognitive Anforderungsdimensionen von Klassenführungsexpertise zu erfassen, wurde ein Testverfahren entwickelt, das anhand von 4 Videovignetten und 27 Items die Genauigkeit der Wahrnehmung (1), die holistische Wahrnehmung (2) und die Rechtfertigung einer Handlung (3) als situationsspezifische Eigenschaften von Klassenführungsexpertise misst. In der vorliegenden Generalisierbarkeitsstudie wurde unter Verwendung einer Stichprobe (n = 188) von Lehramtsstudierenden, Referendaren und Referendarinnen sowie berufstätigen Lehrpersonen den Fragen nachgegangen, (a) wieviel Varianz auf die verschiedenen Facetten (Personen, Videos, Items) zurückzuführen ist sowie b) ob sich die Generalisierbarkeit der Befunde durch eine höhere Anzahl an Videovignetten verbessern lässt. Die Ergebnisse zeigen erwartungskonform, dass der Großteil der erklärbaren Varianz auf die Items (22%) zurückzuführen ist. Die Videovignetten (0.54%) bzw. die Interaktion der Videos mit den Personen (1.77%) erklären hingegen nur einen marginalen Varianzanteil. Es bleibt ein großer Anteil nicht aufzuklärender Residualvarianz (66%). Der Generalisierbarkeitskoeffizient liegt mit $E\rho^2 = .75$ im zufriedenstellenden Bereich und lässt sich durch eine höhere Anzahl an Videos nur geringfügig steigern ($E\rho^2 = .84$ bei 10 Videos). Die Ergebnisse weisen darauf hin, dass die gewählten Videovignetten eine repräsentative Auswahl an Unterrichtssituationen darstellen, eine höhere Anzahl an Videos aus ökonomischen Gründen jedoch nicht zu empfehlen ist.

Schlagwörter: Generalisierbarkeitstheorie, Klassenführung, Lehrerkompetenzen, Videobasiertes Messen

How can classroom management expertise be measured? An examination of a video-based assessment for teachers using generalizability theory

Abstract

Classroom Management Expertise (CME) is a substantial feature of teachers' professional competence. A high level of CME has positive effects on both the quality of teaching and the success of special education interventions. To measure 3 cognitive demands of CME (accuracy of perception, holistic perception, justification of action), a test consisting of 4 video clips and 27

items was developed. In this generalizability study, we used a sample (n = 188) of undergraduate student teachers, graduate student teachers and trained teachers to analyze a) how much of the variance can be attributed to various factors (person, video, item) and b) the number of video clips we need for a good generalizability. Most of the variance is attributable to the items (22%). Only a marginal amount of variance is attributable to the video clips (0.54%) and to the videos' interaction effect with the persons (1.77%). A large amount of residual variance remains unexplainable (66%). The generalizability coefficient reaches an acceptable value of Ep^2 = .75, while there is just a slight increase of the G coefficient when more clips are used (Ep^2 = .84 for 10 clips). Our results underline the generalizability of the video clips used in the test. Thus from a cost-benefit perspective, the use of a larger number video clips is not recommended.

Keywords: Classroom Management, Generalizability Theory, Teacher Competence, Video-based Assessment

Problemstellung

Das deutsche Bildungssystem hat in den letzten Jahren einen Wandel vollzogen, der sich bis heute fortsetzt und in den nächsten Jahren noch anhalten wird. Dafür verantwortlich sind vor allem zwei zentrale Ereignisse: Zum einen durchläuft die Bildungsforschung in Folge der schlechten Ergebnisse Deutschlands im Rahmen von Schulleistungsstudien wie PISA und TIMSS eine empirische Wende. Die damit einhergehende stärkere Kompetenzorientierung bei Schülerinnen und Schülern hat dazu beigetragen, dass a) Bildungspläne grundlegend überarbeitet wurden und b) sowohl Bildungspolitik als auch Bildungsforschung zunehmend datengestützt und evidenzbasiert vorgehen (Bromme, Prenzel & Jäger, 2014). Zum anderen hat sich Deutschland mit der Ratifizierung der UN-Konvention über die Rechte von Menschen mit Behinderungen zur Umsetzung eines inklusiven Schulsystems verpflichtet, was mit schulsystemischen Neuerungen sowie Veränderungen des unterrichtlichen Handelns einhergeht (z.B. Ellinger & Stein, 2012; Lindsay, 2007; Melzer & Hillenbrand, 2015).

Im Zuge dieser Veränderungen ist die Frage danach, was „guten Unterricht" (vgl. Helmke, 2014a, S. 17) ausmacht und wovon er abhängt, wieder zunehmend in den Fokus gerückt. Diese Frage ist jedoch nicht einfach zu beantworten, da Unterricht hochgradig komplex ist und von diversen Faktoren abhängt. Diese Komplexität wird unter anderem dadurch deutlich, dass es verschiedene Forschungsansätze und theoretische Modelle gibt, die sich diesem Thema annehmen und die Qualität von Unterricht zu ergründen versuchen (z.B. Tillmann, 2014). Ein Modell, das die komplexe Wirkungsweise unterrichtlicher Prozesse aufgreift, ist das Angebot-Nutzungs-Modell von Helmke (2014a). Grundgedanke dieser theoretischen Konzeption ist, dass der Unterricht in seiner Gesamtheit ein Angebot umfasst, das nicht zwangsläufig zu wünschenswerten Wirkungen (z.B. fachspezifische und fachübergreifende Kompetenzen) auf Seiten der Schülerinnen und Schüler führt. Vielmehr hängt die Wirksamkeit vom effizienten Zusammenspiel zwischen den angebotenen Lehr- und Lerninhalten sowie deren Nutzung durch die Schülerinnen und Schüler ab.

Die Nutzung des Angebots wird maßgeblich darüber gesteuert, wie es von der Lehrkraft aufbereitet und präsentiert wird (Helmke, 2014a). Dafür wiederum ist spezifisches Wissen und Können erforderlich. Die professionelle Kompetenz von Lehrkräften stellt somit einen zentralen Bedingungsfaktor für die Qualität von Unterricht und damit für die Lernerfolge von Schülerinnen und Schülern dar. Für die professionelle Kompetenz von Lehrkräften hat sich dabei in den letzten Jahren ein Modell etabliert, welches diese Kompetenz in motivationale Orientierungen, die Fähigkeit zur Selbstregulation, Überzeugungen/Werthaltungen

(*beliefs*) sowie professionelles Wissen von Lehrkräften differenziert (Baumert & Kunter, 2006; Blömeke, Kaiser & Lehmann, 2008; Blömeke et al., 2009). Für das professionelle Wissen von Lehrkräften wiederum hat sich eine Segmentierung in Anlehnung an Shulman (1986, 1987) in die drei Bereiche Fachwissen, fachdidaktisches Wissen sowie pädagogisches Wissen als geeignet erwiesen. Letzteres ist hierbei als fachübergreifendes Wissen von Lehrkräften und damit von weitgreifender Wichtigkeit für jede Lehrkraft zu verstehen. Für die Ausgestaltung und den Erfolg von Unterricht jenseits fachspezifischer Faktoren ist das pädagogische Wissen als eine der zentralen kognitiven Komponenten der Lehrerprofessionalität anzusehen (z.B. Baumert & Kunter, 2006). Klassenführung stellt dabei eine dieser substantiellen, fachübergreifenden Anforderungen dar, deren Messbarkeit den Schwerpunkt des vorliegenden Beitrags darstellt. Neben einer Auseinandersetzung mit Klassenführungsexpertise als wesentlichen Bestandteil der professionellen Lehrerkompetenz wird im Folgenden zunächst Klassenführung und deren Relevanz für den schulischen Alltag sowie den sonderpädagogischen Kontext dargestellt. Darauf aufbauend folgt die Frage nach der Messbarkeit von Klassenführungsexpertise, wobei insbesondere die Analyse eines videobasierten Messinstrumentes unter Anwendung der Generalisierbarkeitstheorie im Fokus steht.

Definition von Klassenführung

Um Klassenführungsexpertise von Lehrkräften zu messen, muss zunächst geklärt werden, wie Klassenführung begrifflich gefasst werden kann und welche Kriterien eine effektive Klassenführung ausmachen. Zunächst lässt sich in Forschung und Literatur begrifflich eine breite Vielfalt an verwendeten Termini finden, wobei englischsprachige Begriffe wie *classroom management* (*expertise*) oder auch deutsche Adaptionen, wie *Klassenführung*, *Klassenmanagement*, *Klassenführungskompetenz* oder *Klassenführungsexpertise* zum Teil sowohl synonym als auch inhomogen verwendet werden. Im Folgenden werden die deutschsprachigen Begriffe Klassenführung als Bezeichnung für die *Anforderung* (kongruent zum englischen Classroom Management) sowie Klassenführungsexpertise (kongruent zur englischen Classroom Management Expertise) als Bezeichnung für das *Wissen* über diese Anforderung, beziehungsweise für die *Fähigkeit* von Lehrkräften, dieses Wissen anzuwenden, verwendet.

Im Hinblick auf das Thema Klassenführung kann auf eine breite Basis an empirischer Forschung, seit den Anfängen von Jacob Kounin (1970), verwiesen werden, deren Befunde in der viel beachteten Meta-Analyse von Hattie (2013), den Überblicksarbeiten von Helmke (2014a, 2014b) sowie systematisiert von Evertson und Weinstein (2006) bzw. Emmer und Sabornie (2014) aufgearbeitet wurden. Kounin (2006) fasst unter Klassenführung die Maßnahmen „zur Schaffung einer effektiven schulischen Ökologie, eines effektiven Lernmilieus“ (S. 148) zusammen. In ähnlicher Weise definieren Evertson und Weinstein (2006) den Begriff Klassenführung: „the actions teachers take to create an environment that supports and facilitates both academic and social-emotional learning“ (S. 4). Die Schaffung und Bereitstellung einer Lernumgebung durch die Lehrkraft, in der sowohl das schulische aber auch das sozial-emotionale Lernen der Schülerinnen und Schüler unterstützt und erleichtert wird, die zur Verfügung stehende Lernzeit optimal genutzt wird sowie die Aufrechterhaltung eines möglichst störungsarmen Unterrichts stehen demnach im Mittelpunkt von Klassenführung.

Beiden genannten Definitionen gemein ist das an einen präventiven Grundgedanken gekoppelte grundlegende Prinzip, dass die Lehrkraft klare Abläufe und Routinen ihres Unterrichts im Klassenzimmer etabliert und das Klassengeschehen vorausschauend steuert (Hennemann & Hillenbrand, 2010). Eine solche proaktive Klassenführung be-

steht aus drei zentralen Dimensionen (ebd., S. 256):

1. Der Lehrer antizipiert mögliche Vorkommnisse; er besitzt einen Plan mit konkreten Handlungsmöglichkeiten, um mit unerwarteten Ereignissen umzugehen.
2. Der Lehrer erkennt an: Das Verhalten und das Lernen des Schülers sind untrennbar miteinander verknüpft.
3. Das pädagogische Handeln bezogen auf die Gruppe ist wichtiger als jenes bezogen auf den Einzelnen.

Einem solchen Verständnis folgend, benennen Evertson und Emmer (2009) auf der Grundlage empirischer Forschung insgesamt elf Kriterien effektiver Klassenführung, die sich in neun proaktive (Regeln und Verfahrensweisen planen und unterrichten, Konsequenzen festlegen, Beaufsichtigen und Überwachen, Verantwortlichkeit der Schüler, Unterrichtliche Klarheit, Kooperative Lernformen, Vorbereitung des Klassenraums, Schaffung eines positiven Lernklimas, Vorbereitung des Unterrichts) und zwei reaktive Strategien (Unterbindung unangemessenen Schülerverhaltens, Strategien für potenzielle Probleme) gliedern lassen, die maßgeblich von der Lehrkraft gesteuert und umgesetzt werden können. Gelingt Lehrkräften eine konsequente und vernetzte Anwendung dieser Kriterien, wirkt sich dies in vielfacher Hinsicht positiv auf verschiedene Bereiche des Unterrichts aus (z.B. Helmke, 2014a).

Positive Wirkungen effektiver Klassenführung

Eine effektive Klassenführung weist in zahlreichen nationalen wie internationalen Studien auf verschiedenen Ebenen positive Wirkungen nach. Zum einen führt sie zu einer Steigerung des Lern- und Leistungsniveaus der Schülerinnen und Schüler (z.B. Brophy, 2006; Einsiedler, 1997; Wang, Haertel & Walberg, 1993). Gleichzeitig bewirkt eine effektive Klassenführung eine Reduktion von Unterrichtsstörungen (z.B. Kounin, 2006) und führt zu einer Stressreduktion auf Seiten der Lehrerinnen und Lehrer (z.B. Evertson & Emmer, 2009; König & Rothland, im Druck; Lopez, Pérez & Ochoa, 2008). In diesem Zusammenhang weisen neuere Befunde darauf hin, dass sich die Belastungen beim Berufseinstieg von Lehrkräften (der sogenannte „Praxis-Schock"; Veenman, 1984) durch die Vermittlung von Klassenführungsstrategien signifikant verringern lassen (Dicke, Elling, Schmeck & Leutner, 2015). Weiterhin ist eine erfolgreiche Klassenführung mit einer Verbesserung sozialer Kompetenzen der Schülerinnen und Schüler korreliert (z.B. Durlak, Weissberg, Dymnicki, Taylor & Schellinger, 2011; Wilson, Lipsey & Derzon, 2003). Demnach lässt sich insgesamt feststellen, dass eine effektive Klassenführung maßgeblich zu einer Verbesserung der Unterrichtsqualität beiträgt (König & Pflanzl, Manuskript zur Begutachtung eingereicht) und dementsprechend auch als eine der vielversprechendsten schulischen Maßnahmen gilt (Hattie, 2013).

Auch im sonderpädagogischen Kontext stellt eine erfolgreiche Klassenführung eine zentrale Wirkvariable für den Lern- und Fördererfolg von Kindern und Jugendlichen dar. So kann eine effektive Klassenführung zum einen sozial-emotionale Kompetenzen auf- und zum anderen Verhaltensstörungen abbauen (Hennemann & Hillenbrand, 2010; Hennemann, Hövel, Casale, Hagen & Fitting-Dahlmann, 2015; Reiber & McLaughlin, 2004). Weiterhin gilt eine effektive Klassenführung als eine der vielversprechendsten Methoden zur Reduktion von externalisierenden Verhaltensproblemen, wie z.B. aggressiven oder unaufmerksamen Verhaltens in der Schule (z.B. Hennemann & Hillenbrand, 2010; Reiber & McLaughlin, 2004). Auch für die Förderung von Kindern und Jugendlichen mit Lernstörungen gelten Methoden, die den Kriterien einer effektiven Klassenführung in der Schule entsprechen, gemeinhin als hilfreich (Grünke, 2006). Aus diesen Gründen gilt ei-

ne effektive Klassenführung als eine wichtige Kernkompetenz von Lehrkräften hinsichtlich des Umgangs mit Schülerinnen und Schülern mit zusätzlichem Förderbedarf vor allem im Kontext inklusiver Settings (Oliver & Reschley, 2010).

Die empirisch belegte Wirksamkeit des Konzepts wirft die Frage auf, wodurch eine effektive Klassenführung bedingt wird und wie sich die genannten Kriterien differenziert und effizient umsetzen lassen (z.B. Ophardt & Thiel, 2013). Wie bereits beschrieben, stellt die Lehrkraft die zentrale Wirkvariable bei der Umsetzung von Klassenführungsstrategien dar, womit ihr in diesem Zusammenhang die größte Bedeutung zukommt. Demnach hängt erfolgreiche Klassenführung zunächst von der Kompetenz der Lehrkraft ab, in einer spezifischen Lehr-Lern-Situation entscheidende Aspekte (z.B. Unterrichtsstörungen oder Ablenkung der Schülerinnen und Schüler) wahrzunehmen und zu interpretieren (Kounin, 1970). Damit gemeint sind Kompetenzen zur Schaffung einer produktiven Lernumwelt (z.B. bezogen auf die Beschaffenheit des Klassenraums oder die Herstellung eines positiven sozialen Klimas) sowie die implizite Berücksichtigung der Kriterien und Strategien effektiver Klassenführung (Doyle, 1985). Einem solchen Verständnis folgend, wird die Fähigkeit zur Anwendung dieser Strategien als Klassenführungsexpertise bezeichnet und in verschiedenen theoretischen Ansätzen modelliert.

Klassenführungsexpertise als wesentlicher Bestandteil einer professionellen Lehrerkompetenz

Die Frage danach, was eine gute Lehrerin/einen guten Lehrer ausmacht, beschäftigt die Forschung zum Lehrerinnen- und Lehrerberuf bereits seit einigen Jahrzehnten (Terhardt, Bennewitz & Rothland, 2014). Wie bereits in der Einführung angedeutet, wird in diesem Zusammenhang die Lehrerprofession vielfach als Kompetenz angesehen, die sich zum einen aus fach- und themenbezogenem Wissen, zum anderen auch aus allgemeinem Wissen über pädagogische Konzepte, Prinzipien, Strategien und Techniken zusammensetzt (Blömeke, Gustafsson & Shavelson, 2015; Grossman, 1990; Shulman, 1987). Im Rahmen dieses Verständnisses gilt die Klassenführungsexpertise als zentraler Bestandteil der professionellen Kompetenz von Lehrkräften, da das Konzept universell im unterrichtlichen Kontext anwendbar ist (König & Lebens, 2012). Klassenführung erweist sich somit neben anderen zentralen Merkmalen der Unterrichtqualität als eher fächerübergreifendes Prozessmerkmal.

Diesem Verständnis folgend, konstituiert sich Klassenführungsexpertise zum einen aus dem Wissen über Strategien und Techniken über Klassenführung sowie deren differenzierter Vernetzung untereinander (König & Kramer, 2016). Zum anderen ist es für eine effektive Klassenführung wichtig, dieses Wissen in konkreten Situationen auch anwenden zu können. Häufig wird in diesem Zusammenhang auch von kontextspezifischen kognitiven Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen (Klieme & Leutner, 2006, S. 879), gesprochen. Gerade letzteres ist die entscheidende Teilkompetenz, hinsichtlich derer sich Lehrkräfte in Experten und Novizen unterscheiden lassen. Im Sinne der Expertiseforschung ist die grundlegende Annahme leitend, dass der Erwerb von Klassenführungsexpertise an Berufserfahrung gebunden und somit erlern- und trainierbar ist (z.B. Helmke, 2014a; Sabers, Cushing & Berliner, 1991). Bromme (2008, S. 159) formuliert, „dass die (erfolgreiche) Tätigkeit von Lehrkräften auf Wissen und Können beruht, das in der Ausbildung in theoretischen und praktischen Phasen gewonnen und dann durch die Berufserfahrung weiter entwickelt wurde“.

Zusammenfassend kann also festgehalten werden, dass Klassenführungsexpertise sowohl als situationsspezifische wie auch als ganzheitliche (holistische) Kompetenz

zu verstehen ist und sich mit zunehmender Berufserfahrung differenziert ausbildet (König & Kramer, 2016). Da bislang im deutschsprachigen Raum jedoch ein eher stabiles und analytisches Verständnis von Kompetenz vorherrscht, schlagen Blömeke et al. (2015) eine Erweiterung von Lehrerkompetenzmodellen vor. In ihrem *P(perception)-I(interpretation)-D(decision-making) model of competence transformation* (ebd., S. 7; *Abbildung 1*) wird Kompetenz als Kontinuum mit vielfachen Übergängen postuliert. Zentraler Aspekt dieser Modellierung ist die Transformation von kognitiven (z.B. Wissen über spezifische Klassenführungs-Strategien) und motivational-affektiven (z.B. das Ziel der breiten Aktivierung einer Lerngruppe) Dispositionen über situationsspezifische Fähigkeiten (i.S. kognitiver Informationsverarbeitungsprozesse) hin zur gezeigten Performanz, die sich in beobachtbarem Verhalten niederschlägt. Dieses Modell macht deutlich, dass die entscheidende Variable im Rahmen der Erforschung von Klassenführungsexpertise die Überführung kontextbezogener Kompetenzen in die Performanz in konkreten Situationen ist (König, 2015a).

Diese Transformation stellt hohe Anforderungen an verschiedene kognitive Dimensionen. König und Lebens (2012) konkretisieren die wissensgesteuerte Verarbeitung von Unterricht anhand von drei Dimensionen kognitiver Anforderungen: (1) *die Genauigkeit der Wahrnehmung* (z.B. in Bezug auf bedeutsame Details in einer spezifischen Situation), (2) *die ganzheitliche (holistische) Wahrnehmung* (i.S. einer kontextspezifischen Wahrnehmung dieser Details sowie der darauf aufbauenden Ableitung von Handlungsalternativen) und (3) *die Rechtfertigung einer Handlung* (also die Interpretation einer spezifischen Situation, die über die reine mentale Repräsentation der holistischen Wahrnehmung hinausgeht). Es liegt auf der Hand, dass die Erfassung dieser Dimensionen aufgrund der starken Kontextbezogenheit sowie der Frage nach der manifesten Modellierung des latenten Merkmals große methodische Herausforderungen mit sich bringt (z.B. Blömeke, König, Suhl, Hoth & Döhrmann, 2015).

Videobasierte Messung von Klassenführungsexpertise

Die empirische Erfassung von Klassenführungsexpertise im o.g. Sinne bzw. der postulierten kognitiven Anforderungsdimensionen zur Überführung von Kompetenz in Performanz stellt eine große methodische Herausforderung dar (z.B. König, 2015a;

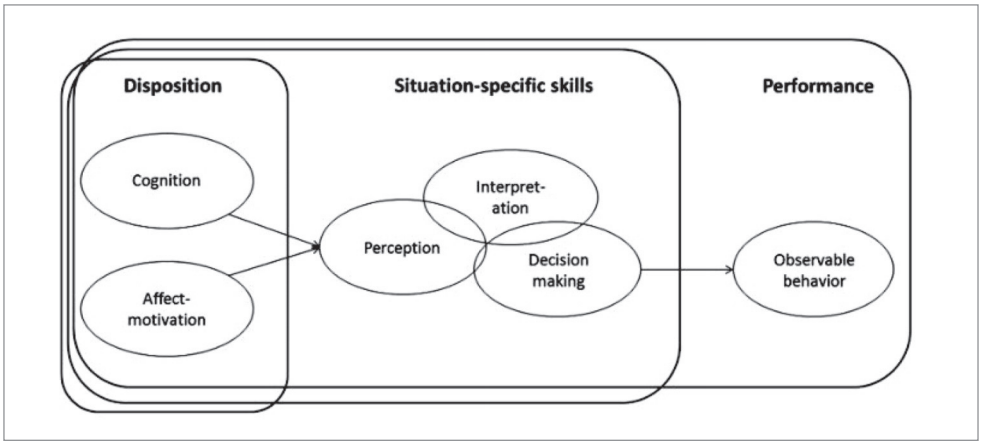

Abbildung 1: Das PID-Modell der Transformation von Kompetenz in Performanz nach Blömeke, Gustafsson & Shavelson (2015, S. 7)

Seidel, Blomberg & Stürmer, 2010). In diesem Kontext ist in der empirischen Bildungsforschung in den letzten Jahren die verstärkte Popularität videobasierter Messinstrumente zu verzeichnen (König, 2015a). Die Erfassung von Klassenführungsexpertise durch videobasierte Instrumente erfolgt in der Regel durch die Nutzung von Videosequenzen konkreter Unterrichtssituationen, in denen Anforderungen an die Klassenführungsexpertise der Lehrkraft gestellt werden und die Anwendung effektiver Klassenführungsstrategien erforderlich ist. Diese Videos werden als *item prompts* eingesetzt, d.h. die Videos sind integraler Bestandteil des Testinstruments, in deren Anschluss die Testpersonen prototypische Testaufgaben beantworten. Videobasierte Testinstrumente weisen in diesem Zusammenhang entscheidende Vorteile gegenüber herkömmlichen Paper-Pencil-Tests auf: Erstens bieten sie die Möglichkeit der situationsspezifischen Kontextualisierung (Blömeke et al., 2015). Zweitens ist mittels videobasierten Testverfahren die holistische Abbildung komplexer Schüler-Lehrer-Interaktionen (i.S. von Klassenführungssituationen) möglich, deren Bewältigung die oben genannten Teilkompetenzen erfordert (König, 2015a). Drittens gelingt die Darstellung der „natürlichen" Klassenraumsituation annäherungsweise besser über Videosequenzen (ebd.). Aus diesen Gründen sind in den letzten Jahren einige dieser Verfahren zur Erfassung der Kompetenzen von Lehrkräften entwickelt und evaluiert worden (z.B. Blömeke et al., 2015; Kersting, Givvin, Sotelo & Stigler, 2010; König & Lebens, 2012; Seidel et al., 2010).

Gleichwohl die Vorteile videobasierter Messungen auf der Hand liegen, bringen sie auch einige methodische Herausforderungen mit sich. (1) So ist aus *ökonomischen Gründen* nur eine begrenzte Anzahl an Videovignetten (als Testaufgaben) nutzbar. Blömeke et al. (2015) weisen darauf hin, dass damit ein Verlust an diagnostischer Genauigkeit und Reliabilität einhergehen kann. (2) In diesem Zusammenhang stellt sich gleichzeitig die Frage der *Repräsentativität* der ausgewählten Videos. Wenn nur wenige Videos genutzt werden können, müssen diese die Grundgesamtheit aller potenziellen Klassenraumsituationen abbilden, damit valide Aussagen über die Kompetenz der Testpersonen zulässig sind. Hier stellt sich die Frage der *Generalisierbarkeit* (Brennan, 2001). (3) Schließlich könnte es noch sein, dass die Items innerhalb eines Itemsets mit dem jeweils vorgeschalteten Video stärker zusammenhängen als mit den übrigen Items des Tests. Allerdings ist die *Homogenität der Items* gerade in einem hierarchischen Kompetenzmodell, wie dem hier postulierten Expertiseansatz, eine wichtige Anforderung an Messinstrumente, da sie die Grundlage zur Differenzierungsfähigkeit hinsichtlich unterschiedlicher Kompetenzstufen ist (Wilbert, 2014).

Methodische Herausforderungen bei der Testentwicklung

Mit der angemessenen Ökonomie unter Erhaltung der Validität und Reliabilität (1), der notwendigen Repräsentativität bzw. Generalisierbarkeit der Videosequenzen (2) und der Homogenität der Items (3) lassen sich drei wesentliche methodische Herausforderungen bei der Entwicklung videobasierter Tests zur Erfassung der Klassenführungsexpertise festhalten. Da sich weder mit Ansätzen der Klassischen Testtheorie (KTT) noch mit denen der Item-Response-Theorie (IRT) alle drei dieser Herausforderungen gleichermaßen forcieren lassen, scheint eine Kombination beider Ansätze sinnvoll (Briggs & Wilson, 2007). Im Kontext der Überprüfung der Ökonomie (1) und Repräsentativität (2) videobasierter Testinstrumente erweist sich der Ansatz der Generalisierbarkeitstheorie (GT) als vielversprechend (z.B. Brennan, 2001; Shavelson & Webb, 1991). Die GT stellt eine Erweiterung der KTT dar, in dem sie den Messfehler nicht als unsystematisch und global annimmt, sondern ihn in seine systematischen Bestandteile zerlegt. Dadurch ist eine simultane Schätzung des An-

teils der verschiedenen Varianzkomponenten an der Gesamtvarianz möglich (*generalizability study; G-Studie*), auf deren Grundlage in einem zweiten Schritt eine Simulation der einzelnen Bedingungen innerhalb einer Varianzkomponente zur Verbesserung der Testgüte (*decision study; D-Studie*) durchgeführt wird. Durch diesen Analyseschritt kann man z.B. überprüfen, wie viele Videovignetten notwendig sind, um valide Ergebnisse generieren zu können. Als Qualitätsmaße werden in der GT zwei Testgüteindizes verwendet: der Generalisierbarkeitskoeffizient $E\rho^2$ wird auf Grundlage der relativen Fehlervarianz berechnet und gibt an, wie generalisierbar die Befunde auf die Gesamtpopulation (in der Terminologie der GT: das *Universum aller zulässigen Bedingungen;* Cronbach, Gleser & Rajaratnam, 1972) sind. Der Zuverlässigkeitskoeffizient Φ (D-Koeffizient) berechnet sich auf Grundlage der absoluten Fehlervarianz und gibt an, wie zuverlässig ein Instrument die Merkmalsausprägung einer Person erfasst (detaillierte Beschreibungen finden sich bei Brennan, 2001 sowie bei Shavelson & Webb, 1991). Für die in diesem Beitrag fokussierte Fragestellung ist demnach nur der G-Koeffizient $E\rho^2$ von Interesse, da in Kombination mit der Varianzaufklärung durch die Videos in Schritt 1 eine Aussage darüber möglich ist, wie zuverlässig das Instrument misst und wie repräsentativ die ausgewählten Unterrichtssequenzen für alle möglichen Unterrichtssituationen sind.

In Bezug auf die Überprüfung der Homogenität der genutzten Items erweist sich die IRT als Mittel der Wahl (z.B. Moosbrugger, 2012; Wilson, 2004). Im Kontext der IRT wird die Schwierigkeit der Items (manifeste Variablen) bezogen auf die tatsächlichen Merkmalsausprägung der Testpersonen (latente Variable) untersucht. Im Falle der videobasierten Messung von Klassenführungsexpertise wäre demnach die Annahme, dass Personen mit einer hohen Klassenführungsexpertise (latentes Merkmal) die spezifischen Items (manifeste Merkmale) ähnlich beantworten, jedoch anders als Personen mit einer geringeren Klassenführungsexpertise. Ist dies der Fall, lässt sich abbilden, wie gut die Items zwischen verschiedenen Kompetenzstufen (in diesem Fall zwischen Novizen und Experten) differenzieren.

Unter Berücksichtigung der genannten methodischen Herausforderungen wurde ein videobasierter Test entwickelt, der die Klassenführungsexpertise von Lehrkräften hinsichtlich der oben genannten kognitiven Anforderungsdimensionen (Genauigkeit der Wahrnehmung, holistische Wahrnehmung, Rechtfertigung einer Handlung) erfasst (König & Kramer, 2016; König & Lebens, 2012). Eine detaillierte Darstellung des Messinstrumentes erfolgt im Methodenteil.

Fragestellung und Hypothesen

Mit Blick auf die formulierten methodischen Herausforderungen bei der Entwicklung und Evaluation videobasierter Testverfahren wurde bei dem hier im Fokus stehenden Instrument in bisherigen Untersuchungen bislang verstärkt den Fragen der Validität und Reliabilität sowie der Homogenität der Testitems nachgegangen (König, 2015b; König & Kramer, 2016; König & Lebens, 2012). Der vorliegende Beitrag legt daher den Schwerpunkt auf die Überprüfung der Generalisierbarkeit und Zuverlässigkeit des Instruments unter Anwendung der Generalisierbarkeitstheorie. Im ersten Schritt wird zunächst betrachtet, wieviel Varianz auf die Unterschiede zwischen den Testpersonen, auf die ausgewählten Videos sowie die ausgewählten Items zurückzuführen ist. Hierbei werden folgende Hypothesen aufgestellt:

H1a: Aufgrund der postulierten unterschiedlichen Klassenführungsexpertise der untersuchten Lehrkräfte, die sich jedoch innerhalb der jeweiligen Personenkategorien (Experten, Novizen, Fortgeschrittene Anfänger) geringfügig unterscheiden sollten, wird davon ausgegangen, dass die Gesamtgruppe der Lehrkräfte (Facette Person) einen moderaten Anteil der Gesamtvarianz aufklärt.

H1b: Die Video-Vignetten (Facette Videos) stellen eine Zufallsauswahl aller möglichen Unterrichtssituationen dar und sollten sich deshalb nicht systematisch hinsichtlich ihrer Bearbeitungskomplexität für die Testpersonen unterscheiden. Daher sollten sie keinen Einfluss auf die Messergebnisse haben und einen geringen Teil der Varianz aufklären.

H1c: Die Items (Facette Items) beziehen sich zum einen auf die jeweiligen Videos und zum anderen auf spezifische kognitive Anforderungsdimensionen. Daher sollten sie stark differenzieren und somit einen großen Teil der Varianz aufklären.

Im zweiten Schritt zielt die zentrale Fragestellung auf die Generalisierbarkeit der Befunde ab und darauf, ob sich die Testgüte des Instruments durch eine veränderte Anzahl an ausgewählten Videosequenzen erhöhen lässt. Es werden folgende Hypothesen aufgestellt:

H2a: Die ausgewählten Unterrichtsvideos sind repräsentativ und damit generalisierbar auf das Universum aller zulässigen Bedingungen. Der G-Koeffizient erreicht daher einen akzeptablen Wert. Auf Grundlage einer Empfehlung von Salvia, Ysseldyke und Bold (2010) sowie der Anwendung in weiteren D-Study-Analysen (z.B. Casale, Hennemann, Volpe, Briesch & Grosche, 2015; Volpe & Briesch, 2012) wird ein Wert von $E\rho^2 \geq .80$ als Kriterium für den G-Koeffizienten gewählt.

H2b: Die Generalisierbarkeit lässt sich durch eine höhere Anzahl an Videosequenzen erhöhen. Daher steigt der G-Koeffizient mit der Anzahl an Videosequenzen.

Methode

Stichprobe

In der vorliegenden Studie werden insgesamt 188 Lehramtsstudierende, Lehramtsanwärter und berufstätige Lehrkräfte aus Köln und der näheren Umgebung untersucht. Die Gesamtstichprobe setzt sich aus 34 Lehrerinnen und Lehrern mit Berufserfahrung (91% weiblich; M = 43.18 Jahre, SD = 9.53; M = 17.9 Jahre Berufserfahrung, SD = 10.4), 40 Lehramtsanwärtern (75% weiblich; M = 27.45 Jahre, SD = 3.46) und 114 Lehramtsstudierenden (79% weiblich; M = 23.21 Jahre, SD = 3.85) zusammen. In Anlehnung an Modelle zur Lehrerexpertise (Berliner, 2001) werden die Personengruppen auf Grundlage der unterschiedlichen Berufserfahrung als Experten, fortgeschrittene Anfänger und Novizen bezeichnet. Die Teilstichprobe der berufstätigen Lehrkräfte setzt sich aus den Gesamtkollegien von zwei allgemeinbildenden Schulen des Kölner Umlands zusammen. Die Personengruppe der Lehramtsanwärterinnen und –anwärter wird aus dem Sample der *Längsschnittlichen Erhebung pädagogischer Kompetenzen von Lehramtsstudierenden und ReferendarInnen* (*LEK-R*) genutzt[1]. Die Gruppe der Lehramtsstudierenden setzt sich aus Teilnehmerinnen und Teilnehmern aus insgesamt drei Didaktik-Seminaren der Universität zu Köln zusammen.

Erhebungsinstrument

Bei dem hier evaluierten Erhebungsinstrument handelt es sich um den von König und Lebens (2012) entwickelten CME-Test (*C*lassroom-*M*anagement-*E*xpertise-Test) zur Erfassung von Klassenführungsexpertise. Wie bereits erläutert, soll der Test testökonomisch einsetzbar sein und im Besonderen das pädagogische Wissen speziell für den Anforderungsbereich der Klassenführung kontextabhängig und handlungsbezogen erfassen.

Insgesamt beinhaltet der CME-Test vier kurze Videosequenzen mit einer Dauer von jeweils 1-2 Minuten mit Unterrichtssituationen, in denen typische Klassenführungssituationen zu sehen sind. Das entscheidende Kriterium für die Auswahl dieser vier Vi-

[1] Das Projekt LEK-R wurde von der Deutschen Forschungsgemeinschaft (DFG) finanziert (Gz.: KO3947/3-2).

deos war, dass authentische und umfassende situative Informationen über Unterrichtssituationen wiedergegeben werden, in denen die Lehrkraft (1) Übergänge gestaltet, (2) Anweisungen (Instruktionen) gibt, (3) mit Schülerverhalten umgeht und (4) mit Schülerrückmeldungen zur Instruktion umgeht (detailliert König & Kramer, 2016). Thematisch stellen zwei Videos die Situation der Instruktion der Schülerinnen und Schüler durch die Lehrkraft in den Mittelpunkt, ein Video beschäftigt sich mit der Situation eines Phasenübergangs und das vierte Video bezieht sich auf die Situation des Schülerfeedbacks im Rahmen der Ergebnissicherung einer Unterrichtsstunde.

Die Videos variieren hinsichtlich der Klassenraumumgebung, d.h. Klasse, Stufe, Fach, Klassenzusammensetzung, Schulform sowie das Alter der Lehrkraft unterscheiden sich in den Situationen. Den Testpersonen werden allerdings keine Informationen über diese Variablen gegeben, um eine systematische Verzerrung der Testwerte zu vermeiden (König & Kramer, 2016)

Im ersten Schritt der Datenerhebung füllen die Testpersonen einen Fragebogen zu verschiedenen soziographischen Angaben (Alter, Geschlecht, Berufserfahrung) aus. Im nächsten Schritt wird der videobasierte Test zur Erfassung der Klassenführungsexpertise eingesetzt. Jeder der vier Videoclips wird nur einmal gezeigt und die Testpersonen dürfen die Items zu den Videos erst nach Anschauen des jeweiligen Videoclips lesen und bearbeiten. Dadurch wird eine standardisierte Anwendung der Videos als *item prompts* sichergestellt. Insgesamt beträgt die Testzeit einschließlich der gezeigten Videos 45 Minuten.

Die im Anschluss an jedes Video schriftlich zu beantwortenden Fragen beziehen sich (a) auf für die Klassenführung wesentliche Spezifika der gesehenen Unterrichtssituation und decken b) die drei verschiedenen kognitiven Anforderungsdimensionen ab (Tabelle 1). Insgesamt werden 24 Fragen an die Befragten gestellt, aus denen in der Auswertung 63 dichotome

Tabelle 1: Aufteilung der genutzten Items pro Video nach Format und Anforderungsdimension

	VC 1					VC 2					VC 3							VC 4						
Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Anforderungsdimension	A	A	J	A	J	A	H	A	A	A	H	A	A	A	J	A	A	A	A	H	H	H	H	J
Format	OR	OR	OR	OR	OR	OR	MC	OR	OR	MC	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	MC	MC	MC	OR

Anmerkungen: VC = Videoclip, A = Accuracy of Perceptiopn (Genauigkeit der Wahrnehmung), H = Holictic Perception (Ganzheitliche Wahrnehmung), J = Justification of Action (Rechtfertigung einer Handlung), OR = open response, MC = multiple choice

Testitems bezogen werden. Für die befragten Testpersonen geht es darum, in den Videos für eine effektive Klassenführung relevante Kriterien zu erkennen bzw. wahrzunehmen. Dabei bezieht sich jeweils eine unterschiedliche Anzahl an Items auf die einzelnen Dimensionen (Genauigkeit der Wahrnehmung: 42 Items, Holistische Wahrnehmung: 11 Items, Rechtfertigung einer Handlung: 10 Items). Weiterhin werden sowohl Multiple-Choice (MC) als auch Open-Response (OR) Item-Formate gewählt. Für die MC-Items werden immer vier Antwortmöglichkeiten zur Verfügung gestellt. Die Antworten auf die OR-Items werden anhand eines umfassenden Manuals kodiert. Dabei sind für jede Testaufgabe Kriterien vorgegeben, die den Erwartungshorizont der Testaufgabe hinreichend widerspiegeln und der Analyse der offenen Antwort zugrunde gelegt werden. Jedes Kriterium wird über eine Variable operationalisiert und mit der Analyse jeder Antwort wird überprüft, ob diese das Kriterium erfüllt (Kodierung = 1) oder nicht (Kodierung = 0).

Mittlerweile liegen bereits einige Überprüfungen zur Testgüte des Instruments vor, die sich vor allem auf die Funktionalität der Items und die Angemessenheit der latenten Konstrukte beziehen. In einer Pilotstudie untersuchten König und Lebens (2012) 19 Lehramtsstudierende im Hauptstudium, 73 Lehramtswärterinnen und –anwärter sowie 16 berufstätige Lehrkräfte mit durchschnittlich ca. 17 Jahren Berufserfahrung. Zur Pilotierung des Instruments wurde den Probandinnen und Probanden eine der Videovignetten (Video „Phasenübergang"; vgl. König & Lebens, 2012, S. 12) gezeigt, an deren Anschluss vier Testfragen (3x OR, 1x MC) beantwortet wurden. Der Fokus der Pilotstudie wurde auf die kognitive Anforderungsdimension *Genauigkeit der Wahrnehmung* gelegt. Die Ergebnisse zeigen, dass mit fortschreitender Expertise auch höhere Lösungshäufigkeiten erzielt werden. Die Skalierung des latenten Merkmals *Genauigkeit der Wahrnehmung* erfolgte mittels IRT-Skalierung im eindimensionalen Raschmodell. Die Befunde deuten darauf hin, dass die Items das Konstrukt homogen messen und somit gut zur Differenzierung über die Personengruppen mit verschiedenen Kompetenzen geeignet sind (alle Items weisen einen *discrimination index* von über .4 auf; ebd., S. 18).

In einer Validierungsstudie mit insgesamt 119 Lehrkräften auf unterschiedlichen Expertisestufen wurde unter anderem der Frage der Reliabilität sowie der Konstruktvalidität des Gesamtinstruments nachgegangen (König, 2015b). Die Overall-Reliabilität liegt bei $\alpha = .70$, was aufgrund der konzeptionellen Anlage des Tests als akzeptabler Wert angenommen werden kann. IRT-Analysen im eindimensionalen Raschmodell zeigen auch hier eine gute Differenzierungsfähigkeit der Items (durchschnittlicher *discrimination index* = .36). Überprüfungen der Konstruktvalidität deuten auf einen stärkeren Zusammenhang mit Skalen, die kognitive Dimensionen messen (z.B. pädagogisches Wissen[2]) als mit denen, die nichtkognitive Dimensionen erfassen (z.B. Lehrer-Burnout), hin (König, 2015b). Weiterhin wird nachgewiesen, dass nur geringe Itemset-Effekte in Bezug auf die spezifischen Videos bestehen (König, 2015b).

Datenanalyse

Zur Überprüfung der o.g. Fragestellungen und Hypothesen wird eine Zwei-Facetten-Generalisierbarkeitsstudie mit geschachtelter Datenstruktur durchgeführt (*p x v x i:v*). Als Differenzierungsfacette werden die untersuchten Lehrkräfte (Facette *p*) gewählt. Als weitere Facetten werden die genutzten Videos (Facette *v*) sowie die Items (Facette *i:v*) festgelegt. Letztere sind innerhalb der Videos geschachtelt, da sich die spezifischen Items jeweils auf die konkreten Videos beziehen. Da jedem Video eine unter-

[2] erhoben in paper-pencil-Form und nicht kontextualisiert zur Klassenführung gemessen

schiedliche Anzahl an Items zugeordnet ist, liegt ein unbalanciertes Design vor.

Die Differenzierungsfacette umfasst insgesamt drei Bedingungen (Experten, Novizen, fortgeschrittene Anfänger). Die insgesamt 188 Personen innerhalb dieser Bedingungen stellen eine Auswahl aus dem *Universum aller möglichen Bedingungen* (i.S. der Gesamtpopulation in der Terminologie der KTT) dar. Die in dieser Studie erzielten Befunde sollen über dieses Universum generalisiert werden, weshalb die Differenzierungsfacette als zufällige Facette definiert wird. Bei den Videos mit den Unterrichtssituationen handelt es sich ebenfalls um eine Auswahl aller möglichen Unterrichtssituationen, daher wird auch diese Facette als zufällige Facette festgelegt. Die Items sind hingegen sehr spezifisch und weisen einen eindeutigen Bezug zu den jeweiligen Videos auf. Daher wird diese Facette als fest definiert.

Die Datenanalyse wird mit der Software urGenova 2.1 (Brennan, 2001), die sich für die Analyse geschachtelter, unbalancierter G-Studien eignet, durchgeführt. In einem ersten Schritt (G-Study) werden die Varianzkomponenten der einzelnen Facetten sowie deren Interaktionen untereinander geschätzt. Auf Grundlage der Befunde aus der G-Studie wird eine D-Study über die Facette Videos durchgeführt, um zu überprüfen, ob sich die Testgüte mit einer modifizierten Anzahl an Videos verbessern lässt. Die Testgüte wird über den Generalisierbarkeitskoeffizienten (G-Koeffizienten) ermittelt. Dieser berechnet sich wie folgt:

$$Ep^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$$

Er setzt sich somit aus der Varianz der universalen Werte der Personen (σ_p^2) in Beziehung zur Summe aus dieser Varianz und der Varianz der gemessenen Personenwerte (σ_δ^2), also der relativen Fehlervarianz, zusammen. In der vorliegenden Simulationsstudie wird somit überprüft, ob sich mit einer unterschiedlichen Anzahl an Videos die relative Fehlervarianz verringern und der G-Koeffizient erhöhen lässt. Er ist also äquivalent zu klassischen Reliabilitätskoeffizienten zu interpretieren (z.B. Brennan, 2001; Eisend, 2007).[3]

Ergebnisse

G-Study

Die Ergebnisse der G-Studie (Tabelle 2) zeigen, dass die Personen knapp zehn Prozent der Gesamtvarianz aufklären (9.69%) (Hypothese H1a). Die Variation in den Testwerten ist also zu einem moderaten Anteil auf die unterschiedliche Ausprägung der Klassenführungsexpertise der Personengruppen zurückzuführen. Die Videos klären hingegen nur einen unbedeutenden Anteil der Gesamtvarianz auf (0.54%), ebenso deren Interaktion mit den Personen (1.77%) (Hypothese H1b). Es gibt also nur einen sehr kleinen, unbedeutenden Einfluss der Videos auf die erzielten Testwerte. Der größte erklärbare Anteil der Gesamtvarianz (22%) ist auf die Facette der Items (geschachtelt innerhalb der Videos) zurückzuführen. Unterschiede in den Testwerten sind also in bedeutsamen Maße auf die Spezifität der Items zurückzuführen (Hypothese H1c). Es bleibt ein recht großer Anteil an nicht aufzuklärender Residualvarianz (66.01%).

D-Study

Die Ergebnisse der D-Studie zeigen, dass der Generalisierbarkeitskoeffizient für das vorliegende Modell mit vier Videos mit $Ep^2 = .75$ im zufriedenstellenden Bereich liegt, wenngleich er unter dem a priori festgelegten kritischen Wert bleibt (Hypothese H2a). Wie erwartet, erhöht sich der G-Koef-

[3] Eine ausführliche Erläuterung des statistischen Modells der GT kann an dieser Stelle aus Platzgründen nicht geleistet werden und ist bei Brennan (2001) sowie Shavelson und Webb (1991) ausführlich nachzulesen.

Tabelle 2: Ergebnisse der Varianzkomponentenschätzung der G-Studie mit p x i:v-Design

Facette	df	T	SS	MS	VC	VC (%)
Person p	187	1678.96	142.46	0.76	0.02	9.69
Video v	3	1571.92	35.42	11.81	0.00	0.54
Item i:v	20	1778.40	206.48	10.32	0.05	22.0
p X v	561	1819.72	105.34	0.19	0.00	1.77
p X i:v, res	3740	2633.00	606.80	0.16	0.16	66.01
$Ep^2 = .75$						

Anmerkungen: df = Freiheitsgrade, T = T-Werte, SS = Quadratsummen, MS = Mittlere Quadratsummen, VC = Varianzkomponentenschätzung

fizient mit der Anzahl der Videos (Hypothese H2b) (Abbildung 2). Bei einer Anzahl von nur zwei Videovignetten liegt der G-Koeffizient bei $Ep^2 = .42$. Der kritische Wert wird bei einer Anzahl von sechs Videovignetten ($Ep^2 = .81$) überschritten. Ab einer Anzahl von zehn Videovignetten steigt der G-Koeffizient nur noch unmerklich ($Ep^2 = .84$).

Diskussion

Zusammenfassend bleibt festzuhalten, dass die Qualität von Unterricht hochgradig komplex ist und von verschiedenen Faktoren abhängt. Das Angebot-Nutzungs-Modell (Helmke, 2014a) oder auch das Expertenparadigma (z.B. Bromme, 2008) greifen diese Komplexität auf und stellen die hohe Bedeutsamkeit der Lehrerkompetenz im Kontext von qualitativ hochwertigem Unterricht heraus. Zunächst einmal unabhängig, weil übergeordnet, von Schulform, Klassenstufe, Größe oder Zusammensetzung einer Schulklasse, stellt das professionelle und erfolgreiche Führen einer Schulklasse, ein zentrales übergreifendes Element und eine Herausforderung für jede Lehrkraft dar. Für die erfolgreiche und professionelle Bewältigung dieser beruflichen Anforderung der Klassenführung ist ein fächerübergreifen-

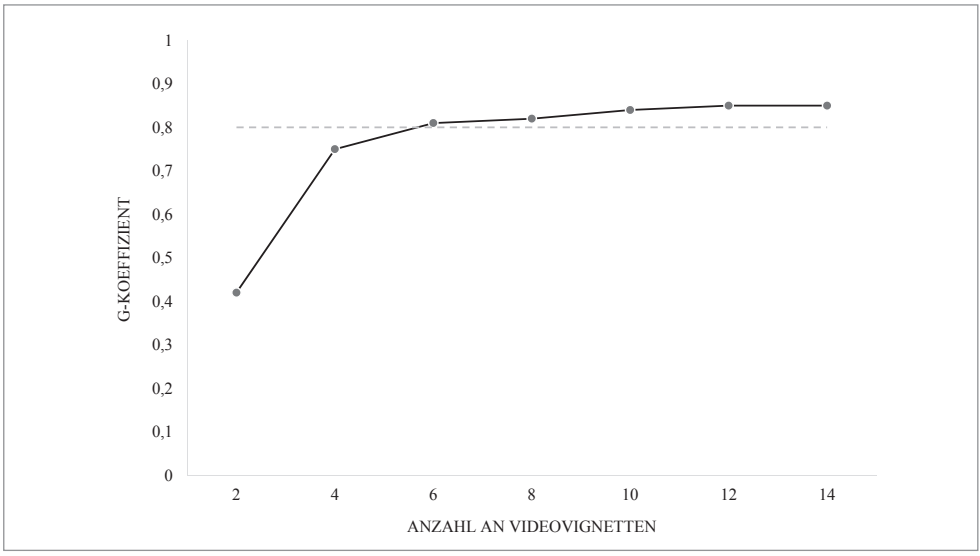

Abbildung 2: Entwicklung des G-Koeffizienten in Abhängigkeit der Anzahl der Videovignetten

des, pädagogisches Wissen notwendig (Doyle 1985, 2006; Evertson & Weinstein, 2006; Helmke, 2014a). In diesem Zusammenhang stellt die Klassenführungsexpertise eine wichtige übergreifende Kompetenz von Lehrkräften dar, deren positive Wirkungen auf verschiedenen Ebenen vielfach nachgewiesen wurden (z.B. Hattie, 2013; Helmke, 2014a). Nicht zuletzt gewinnt Klassenführung auch im Zuge der Umsetzung eines inklusiven Schulsystems und der damit einhergehenden steigenden Heterogenität der Schülerschaft weiter an Relevanz – gerade auch aufgrund der vielfach nachgewiesenen Wirksamkeit im sonderpädagogischen Kontext. Im Rahmen der Messbarkeit von Klassenführungsexpertise wird zunehmend auf videobasierte Testverfahren zurückgegriffen, da diese eine situationsspezifische Kontextualisierung ermöglichen und die „natürliche" Klassenraumsituation annäherungsweise darstellen (König, 2015a). Diese Form der Messung geht allerdings mit einigen methodischen Herausforderung einher, die in diesem Beitrag forciert wurden.

Diskussion hinsichtlich Fragestellung und Hypothesen

Ziel der vorliegenden Studie war es, die Generalisierbarkeit und Zuverlässigkeit des hier vorgestellten videobasierten CME-Tests zur Erfassung der Klassenführungsexpertise zu überprüfen. In diesem Zusammenhang wurde auf die Generalisierbarkeitstheorie zurückgegriffen, um a) den Einfluss verschiedener Fehlerquellen auf die Gesamtvarianz in den Werten simultan zu schätzen und b) durch die Simulation der Bedingungen innerhalb der Facette der Videovignetten eine mögliche Erhöhung der Generalisierbarkeit zu analysieren.

Die Ergebnisse zeigen, dass der Großteil der erklärbaren Varianz auf die Facette der Items rückführbar ist (22%). Dies lässt die Schlussfolgerung zu, dass Unterschiede in den Testwerten zu einem bedeutsamen Anteil auf die Spezifität der Items und deren Kopplung an die Videos zurückzuführen ist. Dieser Befund verwundert nicht, da die Items sich a) sehr spezifisch auf die jeweiligen Situationen in den Videos beziehen, b) unterschiedliche Itemformate pro Video genutzt und c) unterschiedliche Anforderungsdimensionen pro Video geprüft werden. Allerdings konnte in bisherigen Studien Testlet-Effekte der Itemsets bereits negiert werden (König, 2015b). In diesem Kontext scheint die Generalisierbarkeitstheorie prädestiniert dafür, im Rahmen einer experimentellen Überprüfung verschiedener Itemformate in weiteren Studien Anwendung zu finden (z.B. Briesch, Kilgus, Chafouleas, Riley-Tillman & Christ, 2013; Volpe & Briesch, 2012). Auch mit Blick auf die den hohen Anteil nicht aufzuklärender Residualvarianz (66.01%) scheinen anknüpfende Generalisierbarkeitsstudien, die weitere, in der vorliegenden Studie nicht miteinbezogenen Facetten spezifizieren, sinnvoll.

Diese hohe Residualvarianz kann in zweierlei Hinsicht interpretiert und diskutiert werden: Zum einen weist sie darauf hin, dass das Konstrukt der Klassenführungsexpertise nicht ausschließlich aus den mit dem vorliegenden Test gemessenen kognitiven Dimensionen bestehen könnte. Vielmehr scheinen auch nicht-kognitive Merkmale auf das Konstrukt zu wirken. So wurden in einer Studie von König und Rothland (im Druck) Korrelationen des Expertise-Maßes zu non-kognitiven Variablen wie Lehrer-Burnout und Selbstwirksamkeit nachgewiesen. Zum anderen werden in der GT nur bestimmte Facetten, die für die Fragestellung von Interesse sind, in die Modellierung miteinbezogen. So lassen sich diese zwar untersuchen, weitere Facetten, die einen Einfluss auf die Messung haben könnten (z.B. die Testsituation oder der Messzeitpunkt), bleiben dabei jedoch unberücksichtigt (Briesch et al., 2014).

Ein weiterer, erwartungskonformer Befund der Studie ist die moderate Varianzaufklärung durch die Facette der Person (9.69%). Zwar unterscheiden sich die Personengruppen hinsichtlich der Ausprägung

ihrer Klassenführungsexpertise, innerhalb ihrer Gruppen weisen sie jedoch ähnliche Expertisestufen auf. Außerdem lassen sich zwei weitere Interpretationen aus diesem Resultat ableiten: Zum einen sind die Grenzen zwischen Novizen, fortgeschrittenen Anfängern und Experten dimensional und nicht immer trennscharf zu betrachten (z.B. Berliner, 2001). Zum anderen gibt es Anforderungsbereiche und Items (z.B. die Rechtfertigung einer Handlung), in denen keine substanziellen Unterschiede zwischen den Personengruppen feststellbar sind (König & Lebens, 2012; König, 2015b).

Die geringe Varianzaufklärung durch die Videovignetten (0.54%) bzw. deren Interaktion mit den Personen (1.77%) ist ebenfalls ein erwartungskonformer Befund. Die ausgewählten Videos haben kaum Einfluss auf die Unterschiede in den Testwerten, was dahingehend zu interpretieren ist, dass sie sich nicht systematisch hinsichtlich ihrer „Schwierigkeit", wesentliche Kriterien in der Situation wahrzunehmen, unterscheiden. Die oben genannten Vorteile videobasierter Messung zur Erfassung kontextualisierter Kompetenzen scheinen also auch in der Empirie ihren Niederschlag zu finden, wenngleich auch hier ein experimenteller direkter Vergleich der gemessenen abhängigen Variablen mit einem Paper-Pencil-Test wünschenswert wäre (König, 2015a).

In diesem Zusammenhang wurde der Frage der Generalisierbarkeit der Ergebnisse sowie ihrer Verbesserung durch eine Veränderung der Facettenstufen nachgegangen. Der G-Koeffizient liegt bei $Ep^2 = .75$ und damit knapp unter dem festgelegten kritischen Wert (Salvia et al., 2010). Aus zweierlei Gründen kann dieser Befund dennoch als zufriedenstellend angenommen werden: Erstens wird der kritische Wert bereits ab einer Anzahl von sechs Videovignetten überschritten ($Ep^2 = .81$) und steigt ab dann mit einer steigenden Anzahl an Videos nur noch marginal an. Eine erhöhte Anzahl an Videos, die zwar einen kleinen Anstieg der Generalisierbarkeit, aber einen große Beeinträchtigung der Ökonomie des Instruments mit sich bringt, erscheint vor diesem Hintergrund wenig sinnvoll. Zweitens wurde der kritische Wert auf Grundlage von Empfehlungen aus bisherigen G-Studien festgelegt, die sich schwerpunktmäßig auf Verhalten von Schülerinnen und Schülern konzentrieren (Casale et al., 2015; Volpe & Briesch, 2012). Dieser Wert wird recht hoch angelegt, da die entwickelten Instrumente als wichtige Grundlage für Entscheidungen über konkrete Förderangebote für Kinder und Jugendliche dienen sollen und sie daher extrem hohen Anforderungen entsprechen müssen (z.B. Casale, Hennemann, Huber & Grosche, 2015). Es bleibt offen, ob diese hohen Standards für das hier überprüfte Instrument angemessen sind oder ob auch ein G-Koeffizient ab $Ep^2 = .70$, wie er in diversen, nicht anwendungsorientierten Studien festgelegt wird (z.B. Pietsch & Tosana, 2008; Praetorius, 2014), bereits als akzeptabel gelten kann.

Inhaltliche Limitationen

Betrachtet man die in dieser Studie berichteten Befunde in Kombination mit den bisherigen testtheoretischen Absicherungen des Verfahrens (v.a. König, 2015b; König & Kramer, 2016; König & Lebens, 2012), kann konstatiert werden, dass die drei kognitiven Anforderungsdimensionen von Klassenführungsexpertise valide und reliabel erfasst werden können. Dennoch stellt sich die Frage, welchen Beitrag videobasierte Messinstrumente generell zur Erfassung von Lehrerkompetenz leisten können, wenn man sich an einem kontinuierlichen Kompetenzverständnis – wie im PID-Modell (Abbildung 1; Blömeke et al., 2015) postuliert – orientiert.

Das in diesem Beitrag überprüfte Verfahren legt den Schwerpunkt auf die Wahrnehmung spezifischer Kriterien in einer konkreten Unterrichtssituation, also auf die situationsspezifischen Eigenschaften der Lehrkräfte. Damit wird ein wesentlicher Bereich im Rahmen der Transformation von kognitiven Dispositionen in Performanz ab-

gedeckt (Blömeke et al., 2015). Gleichwohl diese Eigenschaften eine wesentliche Voraussetzung für die Performanz, also das „Zeigen" der Kompetenz, sind, kann nicht monokausal davon ausgegangen werden, dass eine hohe Expertise in diesem Bereich mit einer hohen Performanz einhergeht, da sich auf Video aufgezeichnete Unterrichtssituationen von realen Unterrichtssituationen unterscheiden (Kaiser, Busse, Hoth, König & Blömeke, 2015). Eine Möglichkeit zur Erfassung der Performanz können daher systematische Verhaltensbeobachtungen im Rahmen von In-Vivo-Experimenten sein, wie sie in den letzten Jahren beispielsweise durch die Videographie des Unterrichts oder die Unterrichtsbeobachtung durch geschulte Beobachterinnen und Beobachter umgesetzt wird. Der Vorteil dieser methodischen Herangehensweise ist, dass die Erfassung der Performanz (in Form beobachtbaren Verhaltens) so nah wie möglich am tatsächlichen Auftreten erfolgt und die Güte der Messung somit hoch reliabel und valide ist (Cone, 1978). Nachteile dieser Vorgehensweise sind zum einen, dass mit der beobachtenden Person eine Störvariable im natürlichen Umfeld implementiert wird (Schmidt-Atzert & Amelang, 2012). Zum anderen sind in-vivo-Experimente sowohl finanziell als auch personell sehr aufwändig und damit nicht mehr ökonomisch. Darüber hinaus besteht die Herausforderung, die erhobenen Messungen, beispielsweise in Form von Unterrichtsvideographie oder –beobachtungen, zu operationalisieren und in Daten zu „übersetzen", um sie mess- und vergleichbar zu machen. Diese Transformation, aber auch die Erfassung der Daten (zum Beispiel in Form von Ratings) wird durch zahlreiche Faktoren beeinflusst, welche die Aussagekraft videobasierter Unterrichtsforschung einschränken können (z.B. Praetorius, 2014). Hierzu gehört etwa die Identifikation geeigneter Raterinnen und Rater oder die Art des Ratings (z.B. hochinferent vs. niedrig inferent oder quantifizierendes vs. qualitatives Vorgehen), weshalb in-vivo-Experimente nicht per se als reliabler und valider angesehen werden können.

Schließlich sei noch angemerkt, dass in der vorliegenden Studie nur Regelschullehrkräfte (bzw. Studierende des Regelschullehramts) untersucht wurden. Ebenfalls stammen die in den Videovignetten präsentierten Unterrichtssituationen aus der allgemeinen Schule. Dennoch haben die in dieser Studie erzielten Befunde hohe fachliche Relevanz für die sonderpädagogische Praxis. Zum einen sind die positiven Wirkungen einer effektiven Klassenführung im sonderpädagogischen Kontext (s.o.) zu einem großen Maß auf die Klassenführungsexpertise von Lehrkräften zurückzuführen. Diese stellt somit eine der Grundlagen präventiven Handelns im heterogenen Klassenraum, insbesondere im Umgang mit Lern-, Verhaltens- und Disziplinproblemen dar (z.B. Oliver & Reschley, 2010). Zum anderen beziehen sich die Videovignetten zwar auf Unterrichtssituationen aus allgemeinen Schulen, vor dem Hintergrund der auch in diesen Schulen hohen Prävalenzraten von z.B. Lern- und Verhaltensproblemen sowie dem zunehmenden Wandel zu inklusiven schulischen Lerngruppen bilden sie dennoch ein für die sonderpädagogische Profession hoch relevantes Feld ab. Dennoch wäre zu überlegen, inwiefern eine zusätzliche Validierung des Instruments unter Berücksichtigung sonderpädagogischer Spezifika, d.h. Untersuchung einer Validierungsstichprobe aus Förderschullehrkräften sowie Nutzung von Videovignetten, die Unterrichtssituationen mit Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf einbeziehen, sinnvoll wären. Zumindest aus testtheoretischer Sicht könnten hier Unterschiede hinsichtlich der Funktionalität der Items sowie der Inhaltsvalidität der Videos vermutet werden (Bühner, 2011).

Methodische Einschränkungen

Weiterhin weist die vorliegende Studie methodische Einschränkungen auf, die es zu diskutieren gilt. Die Generalisierbarkeits-

theorie bietet im Kontext der Testentwicklung zwar einige Vorteile, wie z.B. die Möglichkeit verschiedene Fehlerquellen simultan zu schätzen und die Erhöhung der Testgüte auf Grundlage dieser Schätzungen zu simulieren (Brennan, 2001). Hieraus lassen sich wertvolle Schlüsse ziehen, die auch im Rahmen der Entwicklung des hier überprüften Instruments wertvolle Implikationen liefern können. Gleichzeitig ist jedoch auf problematische Aspekte der methodischen Vorgehensweise hinzuweisen. So besteht im Rahmen von G-Studien die Gefahr negativer Varianzkomponentenschätzungen (Eisend, 2007). Schätzfehler dieser Art resultieren meist aus Stichproben, die zu klein für Varianzanalysen sind. Die Frage nach der für G-Studien richtigen Stichprobengröße ist allerdings noch nicht vollends geklärt (Briesch, Swaminathan, Welsh & Chafouleas, 2014). Hilfreich scheint eine Empfehlung von Webb, Rowley und Shavelson (1988), wonach mindestens 20 Personen und mindestens zwei Bedingungen pro Facette erforderlich sind. Demnach wäre die Stichprobengröße von n = 188 (114, Lehramtsstudierenden, 40 Lehramtsanwärter innen und Lehramtsanwärterinnen, 34 Lehrkräften mit Berufserfahrung) in der hier vorliegenden Studie als angemessen zu betrachten.

Ein weiterer Kritikpunkt, der im Kontext von GT-Analysen thematisiert wird, ist die Aussagekraft der Zuverlässigkeit der Ergebnisse über die gewählten Bedingungen der Stichprobe hinaus. In diesem Zusammenhang sind zwei Aspekte zu nennen: Zum einen ist die GT eine „Stichprobentheorie", deren ausgewählte Bedingungen immer vom Erkenntnisinteresse gelenkt werden (Shavelson & Webb, 1991). Daher sind die Aussagen auch nur in diesem Kontext bzw. in Bezug auf das jeweilige Universum zulässiger Bedingungen generalisierbar. Die Zuverlässigkeit der Befunde kann jedoch entweder über die mehrfache Durchführung von G-Studien mit unterschiedlichen Schwerpunkten oder über die Kombination mit anderen testtheoretischen Ansätzen gesteigert werden (Briggs & Wilson, 2007). Zum anderen geht die vorliegende Studie nur von einem stark umgrenzten Bereich im Rahmen der Entwicklung eines videobasierten Instruments zur Erfassung der Klassenführungsexpertise aus. Differenzierte Item- und Reliabilitätsanalyse können und sollen in diesem Kontext nicht geleistet werden, sondern wurden vielmehr bereits in anderen Studien durchgeführt (König, 2015b; König & Kramer, 2016; König & Lebens, 2012). Vor diesem Hintergrund stellen die hier erzielten Befunde eine wertvolle Ergänzung der bisherigen Forschung zu diesem Instrument dar.

Literaturverzeichnis

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften., *Zeitschrift für Erziehungswissenschaft, 9,* 469-520.

Berliner, D.C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463-482.

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. (2015). Beyond dichotomies: Viewing competence as a continuum. *Zeitschrift für Psychologie, 223*, 3-13.

Blömeke, S., Kaiser, G., & Lehmann, R. (Hrsg.) (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Münster: Waxmann.

Blömeke, S., Kaiser, G., Lehmann, R., König, J., Döhrmann, M., Buchholtz, C. & Hacke, S. (2009). TEDS-M: Messung von Lehrerkompetenzen im internationalen Vergleich. In R. Mulder, O. Zlatkin-Troitschanskaia, K. Beck, N. Reinhold & D. Sembill (Hrsg.), *Professionalität von Lehrenden – Zum Stand der Forschung*. Weinheim: Beck, S. 181-210.

Blömeke, S., König, J., Suhl, U., Hoth, J. & Döhrmann, M. (2015). Wie situationsbezogen ist die Kompetenz von Lehrkräften? Zur Generalisierbarkeit der Ergebnisse von videobasierten Performanztests. *Zeitschrift für Pädagogik, 61,* 310-327.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Briesch, A. M., Kilgus, S.P., Chafouleas, S.M., Riley-Tillman, T.C. & Christ, T.J. (2013). The Influence of Alternative Scale Formats on the Generalizability of Data Obtained From Direct Behavior Rating Single-Item Scales (DBR-SIS). *Assessment for Effective Intervention, 38,* 127-133.

Briesch, A. M., Swaminathan, H., Welsh, M. & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology 52,* 13-35.

Briggs, D. C. & Wilson, M. (2007). Generalizability in Item Response Modeling. *Journal of Educational Measurement, 44,* 131-155.

Bromme, R. (2008). Lehrerexpertise. In W. Schneider & M. Hasselhorn (Hg.). Handbuch der Pädagogischen Psychologie. Göttingen: Hogrefe.

Bromme, R., Prenzel, M. & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. Eine Analyse von Anforderungen an die Darstellung, Interpretation und Rezeption empirischer Befunde. *Zeitschrift für Erziehungswissenschaft, 17*, 3-54.

Brophy, J. (2006). Observational Research on Generic Aspects of Classroom Teaching. In P.A. Alexander & P.H. Winne (Hrsg.). *Handbook of Educational Psychology*, (755-780). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

Casale, G., Hennemann, T., Huber,C. & Grosche, M. (2015). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. *Heilpädagogische Forschung, 41*, 37-54.

Casale, G., Hennemann, T., Volpe, R.J., Briesch, A.M. & Grosche, M. (2015). Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen des Lern- und Arbeitsverhaltens in einer inklusiven Grundschulklasse. *Empirische Sonderpädagogik, 7*, 258-268.

Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy, 9,* 882-888.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements. Theory of Generalizability for scores and profiles.* New York: John Wiley & Sons.

Dicke, T., Elling, J., Schmeck, A. & Leutner, D. (2015). Reducing reality shock: The effects of classroom management skolls training on beginning teachers. *Teaching and Teacher Education, 48*, 1-12.

Doyle, W. (1985). Recent Research on Classroom Management: implications for Teacher Preparation. *Journal of Teacher Education, 36*, 31-35.

Doyle, W. (2006). Ecological approaches to classroom management. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 97-125) Mahwah, NJ: Erlbaum.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D. & Schellinger, K. B. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development, 82*, 405-432.

Einsiedler, W. (1997). Empirische Grundschulforschung im deutschsprachigen Raum: Trends und Defizite. *Unterrichtswissenschaft, 25*, 291-315.

Eisend, M. (2007). *Methodische Grundlagen und Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung.* Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der

Freien Universität Berlin, N0. 2007/4, ISBN 3938369523.

Ellinger, S. & Stein, R. (2012). Effekte inklusiver Beschulung: Forschungsstand im Förderschwerpunkt emotionale und soziale Entwicklung. *Empirische Sonderpädagogik, 4*, 85-109.

Emmer, E. T. & Sabornie, E. J. (2014). *Handbook of Classroom Management. 2nd Edition*. New York: Routledge.

Evertson, C. M. & Emmer, E. T. (2009). *Classroom Management for Elementary Teaches. 8th Edition*. New Jersey: Pearson Education.

Evertson, C. M. & Weinstein, C. S. (2006). *Handbook of Classroom Management. Research, Practice, and Contemporary Issues*. Mahwah, NJ: Lawrence Erlbaum.

Grossman, P. L. (1990), *The making of a teacher: Teacher knowledge and teacher education*. New York: Teachers College Press.

Grünke, M. (2006). Zur Effektivität von Fördermethoden bei Kindern und Jugendlichen mit Lernstörungen. Eine Synopse vorliegender Metaanalysen. *Kindheit und Entwicklung, 15,* 239-254.

Hattie, J. (2013). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. New York: Routledge.

Helmke, A. (2014a). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.

Helmke, A. (2014b). Forschung zur Lernwirksamkeit des Lehrerhandelns. In E. Terhardt, H. Bennewitz & M. Rothland (Hrsg.). *Handbuch der Forschung zum Lehrerberuf, 2. Aufl.*, 807-821. Münster: Waxmann.

Hennemann, T. & Hillenbrand, C. (2010). Klassenführung – Classroom Management. In B. Hartke & K. Koch (Hrsg.), *Förderung in der schulischen Eingangsphase*, 255-279. Stuttgart: Kohlhammer.

Hennemann, T., Hövel, D., Casale, G., Hagen, T. & Fitting-Dahlmann, K. (2015*). Schulische Prävention im Bereich Verhalten*. Stuttgart: Kohlhammer.

Kaiser, G., Busse, A., Hoth, J., König, J. & Blömeke, S. (2015). About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Mathematics Education, 13*, 369-387.

Kersting, N. B., Givvin, K., Sotelo, F. & Stigler, J. W. (2010). Teacher's analysis of classroom video predicts student learning of mathematics: Further exploration of a novel measure of teacher knowledge. *Journal of Teacher Education, 61*, 172-181.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Pädagogik, 52*, 876-903.

König, J. (2015a). Kontextualisierte Erfassung von Lehrerkompetenzen. Einführung in den Thementeil. *Zeitschrift für Pädagogik, 61*, 305-309.

König, J. (2015b). Measuring Classroom Management Expertise (CME) of Teachers: A Video-Based Assessment Approach and Statistical Results. *Cogent Education, 2*(1), 1-15.

König, J. & Kramer, C. (2016). Teacher professional knowledge and classroom management: On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM - The International Journal on Mathematics Education, 48(1),* 139-151.

König, J. & Lebens, M. (2012). Classroom Management Expertise (CME) von Lehrkräften messen: Überlegungen zur Testung mithilfe von Videovignetten und erste empirische Befunde. *Lehrerbildung auf dem Prüfstand, 5(1)*, 3-29.

König, J.; Pflanzl, B. (2016) (Manuskript zur Begutachtung eingereicht). Is teacher knowledge associated with teacher performance? On the relationship between teachers´ general pedagogical knowledge and instructional quality. *European Journal of Teacher Education*.

König, J. & Rothland, M. (2016) (in Druck). Klassenführungswissen als Ressource der Burnout-Prävention? Zum Nutzen von pädagogisch-psychologischem Wissen im Lehrerberuf. *Unterrichtswissenschaft*.

Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rineheart and Winston.

Kounin, J. S. (2006). *Techniken der Klassenführung, 2. Aufl*. Münster: Waxmann.

Lindsay, G. (2007). Educational psychology and the effectiveness of inclusive education/ mainstreaming. *British Journal of Educational Psychology, 77*, 1-24.

Lopez, E.E., Pérez, S.M. & Ochoa, G.M. (2008). Psychosocial adjustment in aggressors, pure victims and aggressive victims at school. *European Journal of Education and Psychology, 1*, 29-39.

Melzer, C. & Hillenbrand, C. (2015). Aufgabenprofile - Welche Aufgaben bewältigen sonderpädagogische Lehrkräfte in verschiedenen schulischen Tätigkeitsfeldern. *Zeitschrift für Heilpädagogik, 66,* 230-242.

Moosbrugger, H. (2012). Item-Response-Theorie. In H. Moosbrugger & A. Kelava (Hrsg.). Testtheorie und Fragebogenkonstruktion, 228-275. Berlin, Heidelberg: Springer.

Oliver, R.M. & Reschley, D.J. (2010). Special Education Teacher Preparation in Classroom Management;: Implication for Students With Emotional and Behavioral Disorders. *Behavioral Disorders, 35,* 188-199.

Ophardt, D. & Thiel, F. (2013). *Klassenmanagement. Ein Handbuch für Studium und Praxis*. Stuttgart: Kohlhammer.

Pietsch, M. & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft, 11*, 430-452.

Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.

Reiber, C. & McLaughlin, T.F. (2004). Classroom Interventions: Methods to Improve Academic Performance and Classroom Behavior for Students with Attention-Deficit/Hyperacitivity Disorder. *International Journal of Special Education, 19 (1),* 1-13.

Sabers, D. S., Cushing, K. S. & Berliner, D. C. (1991). Differences Among Teachers in a Task Characterized by Simultaneity, Multidimensionality, and Immediacy. *American Educational Research Journal, 28(1),* 63-88.

Salvia, J., Ysseldyke, J.E. & Bolt, S. (2010). *Assessment in special and inclusive education, 11th Edition*. Boston, MA: Houghton Mifflin.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik, 5. Aufl*. Heidelberg: Springer.

Seidel, T., Blomberg, G. & Stürmer, K. (2010). „Observer" – Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. *Zeitschrift für Pädagogik, 56. Beiheft*, 296-306.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability Theory. A Primer*. Newbury Park, CA: Sage.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15 (2),* 4-14.

Shulman, L.S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Research, 57*, 1-22.

Terhardt, E., Bennewitz, H. & Rothland, M. (2014). *Handbuch der Forschung zum Lehrerberuf, 2. Aufl*.. Münster: Waxmann.

Tillmann, K. J. (2014). Konzepte der Forschung zum Lehrerberuf. In E. Terhardt, H. Bennewitz & M. Rothland (Hrsg.). *Handbuch der Forschung zum Lehrerberuf, 2. Aufl.*, 308-318. Münster: Waxmann.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*, 143-178.

Volpe, R. J. & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41*, 246-261.

Wang, M. C., Haertel, G. D. & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of educational research*, *63*, 249-294.

Webb, N. N., Rowley, G. L. & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development 21*, 81-90.

Wilbert, J. (2014). Instrumente zur Lernverlaufsdiagnostik: Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik,* 281-308. Göttingen: Hogrefe.

Wilson, M. (2004*). Constructing Measures. An Item Response Modeling Approcach*. New York: Psychology Press.

Wilson, S. J., Lipsey, M. W. & Derzon, J. H. (2003). The Effects of School-Based Intervention Programs on Aggressive Behavior: A Meta-Analysis. *Journal of Consulting and Clinical Psychology, 71*, 136-149.

Gino Casale
Universität zu Köln
Department Heilpädagogik und Rehabilitation
Klosterstraße 79 c
50931 Köln
gino.casale@uni-koeln.de

Erstmalig eingereicht: 15.10.2015
Überarbeitung eingereicht: 13.01.2016
Angenommen: 20.01.2016