Hauke Bartels Universität Bremen
Christoph Kulgemeyer

Ein Videovignettentest zur Messung der Erklärfähigkeit von Lehrkräften

In diesem Beitrag wird ein Online-Testinstrument für die Erklärfähigkeit von Physiklehrkräften vorgestellt, das zwei Ziele verfolgt: (1) die authentische Simulation einer Handlungssituation (das Erklären) sowie (2) eine höhere Testökonomie als bisherige Testformate. Dazu werden zweistufige Items mit handlungsnahen Videovignetten genutzt und – wie in einer realen Unterrichtssituation – Handeln unter Druck durch zeitlich befristete Antwortmöglichkeiten simuliert. Ziel ist es, mit dem Instrument Large-Scale Erhebungen durchführen zu können und dennoch eine hohe prognostische Validität hinsichtlich tatsächlichen Unterrichtshandelns zu gewährleisten. Darüber hinaus sollen Strategien bei der Vorgehensweise des Erklärens sichtbar gemacht werden.

Messung von Erklärfähigkeit als Teil des professionellen Lehrerhandelns

Die Fähigkeit, Schülerinnen und Schülern naturwissenschaftliche Sachverhalte erklären zu können, wird als eine wichtige (z. B. Wilson & Mant 2011) und anspruchsvolle Fähigkeit (z. B. Merzyn 2005) von Lehrkräften beschrieben. Wesentliche Aspekte einer guten Erklärung sind (a) Sachgerechtheit (fachliche Vollständigkeit und Korrektheit) und (b) Adressatengemäßheit (Adaption an den Bedürfnissen von Adressaten) (Kulgemeyer & Schecker 2013).
Wer diese handlungsnahe Fähigkeit mit einem Testinstrument erheben möchte, steht vor einer Herausforderung, die typisch für das Messen handlungsnaher Fähigkeiten von Lehrkräften ist: ein Abwägen zwischen Aufwand und Authentizität der Testung. Das Spektrum möglicher Testformate reicht dabei von Paper-and-Pencil Tests bis zur Videografie von realem Unterricht. Aus testökonomischer Sicht sind schriftliche Tests das Mittel der Wahl, da sie vergleichsweise einfach und objektiv durchgeführt werden können, eine hohe Reliabilität versprechen und sich oft sogar automatisch auswerten lassen. Für videobasierte Studien müssen neben der aufwändigeren Datenerhebung zunächst ein Kodiermanual entwickelt und Rater ausgebildet werden, was einen erheblichen Aufwand mit sich bringt. Mit Blick auf die Authentizität der Handlungen (also die Nähe zu realem Unterrichten), welche die Probanden im Test durchführen müssen, ist das Videografieren von Unterricht jedoch die optimale Lösung. Daher raten viele Autoren für die Messung professionellen Handelns zu Instrumenten, welche zumindest über schriftliche Formate hinausgehen (z. B. Aufschnaiter & Blömeke 2010). Auf der anderen Seite sind zu erwartende Effektstärken bei der Untersuchung von Interventionen in der Lehrerbildung gering und machen damit große Stichproben erforderlich. Eine typische Effektstärke von $d = 0{,}11$ (Hattie 2009, S. 110) erfordert beispielsweise eine Stichprobe von $N = 505$ Probanden.

Performanz-orientiertes Testen

Ein Ansatz zum Umgang mit dieser Schwierigkeit sind sogenannte Performanztests, bei denen professionelles Handeln in standardisierten aber authentischen Situationen erhoben wird (Miller 1990). Beispiele dafür gibt es in vielen Disziplinen, z.B. der Ärzteausbildung (z. B. Walters, Osborn & Raven 2005), der Pilotenausbildung (z. B. Winter, Dodou & Mulder 2012) oder in verschiedenen beruflichen Ausbildungsgängen (Beck, Landenberger & Oser 2016).
In der Domäne der Lehrerbildungsforschung haben Kulgemeyer & Tomczyszyn (2015) einen Performanztest für die Erklärfähigkeit von Physiklehrkräften entwickelt. In diesem Test werden die Teilnehmenden einzeln gebeten, einem Schüler einen physikalischen Sachverhalt zu erklären. Der Schüler ist darauf trainiert, standardisierte Nachfragen zu stellen, die auf die vier Aspekte des Erklärens nach Kulgemeyer & Schecker (2013) abzielen (Sprachniveau,

Mathematisierungsgrad, adäquate Beispiele und Darstellungsformen). Auch wenn hier kein Klassenraumsetting, sondern eher ein Dialog im Sinne einer Nachhilfestunde abgebildet wird, hat dieses Verfahren einen hohen Grad an Authentizität hinsichtlich tatsächlichen Lehrerhandelns. Für die Auswertung ist allerdings eine Kodierung des Videomaterials notwendig, was trotz einer im Vergleich zur Beobachtung von realem Unterricht deutlich reduzierten Anzahl an Freiheitsgraden sehr zeitaufwändig ist (Kulgemeyer & Tomczyszyn 2015).

Ein handlungsnaher Videovignettentest mit geschlossenen, zweistufigen Items
Das erstellte Instrument umfasst zwei Tests, die unterschiedliche physikalische Sachverhalte thematisieren (Impulserhaltung und 3. Newton'sches Axiom) und auf Deutsch und Englisch verfügbar sind. Sie bestehen jeweils aus zwei Teilen. Der erste Teil erhebt Aspekte der Sachgerechtheit und überprüft somit, ob Probandinnen und Probanden in der Lage sind, eine fachlich vollständige und korrekte Erklärung zu geben. Dieser Teil wird mit geschlossenen Single-Select Items realisiert (etwa 15 Minuten). Der zweite Teil erhebt Aspekte des adressatengemäßen Erklärens (etwa 45 Minuten). Hier wird die Fähigkeit getestet, das Erklären an die Bedürfnisse der Adressaten zu adaptieren. Außerdem werden Strategien bei der Auswahl des Vorgehens sichtbar gemacht. Dazu werden zweistufige Items eingesetzt (vgl. z. B. Haagen-Schützenhöfer & Hopf 2014). Diese Items beginnen jeweils mit einer Videosequenz, in der die Lehrkraft einer Schülerin einen Teilaspekt des physikalischen Sachverhalts erklärt. Mit einer Nachfrage der Schülerin stoppt die Sequenz und der Proband muss in der ersten Stufe aus vier vorgegebenen Antworten diejenige auswählen, von der er glaubt, dass sie die beste Möglichkeit zum Fortfahren ist. Ein Zeitlimit zum Treffen der Entscheidung soll den Handlungsdruck einer realen Situation simulieren (Rehm & Bölsterli 2014). In der zweiten Stufe wird er dann gebeten, seine Auswahl zu begründen. Dafür kann aus vorgegebenen Begründungen gewählt werden, die bei der Testentwicklung aus Rückmeldungen im Freitext synthetisiert wurden. Anschließend wird der Dialog zwischen Lehrkraft und Schülerin im nächsten Video fortgeführt. Um eine hohe Authentizität sicherzustellen, wurden weitere Maßnahmen ergriffen. So handelt es sich sowohl bei den gezeigten Dialogen, als auch bei den Auswahlmöglichkeiten um Erklärungen, die bei Probanden beobachtet werden konnten, die am oben beschriebenen Performanztest teilgenommen haben. Die verschiedenen Items decken alle Aspekte des Erklärens nach Kulgemeyer (2013) ab.

Im Teil zum adressatengemäßen Erklären sind alle Antwortmöglichkeiten zur Fortführung der Videovignetten fachlich korrekt. Hauptaspekte bei der Festlegung der besten Antwort waren Hinweise aus der Literatur und die Äußerungen der Schülerin in den Videovignetten. In einem iterativen Prozess wurden die vier Antwortmöglichkeiten der ersten Stufe jedes Items anschließend von einer Expertengruppe (N=10, Physikdidaktiker) diskutiert und überarbeitet, bis ein Konsens über die beste Antwort gefunden wurde. In einer anschließenden Think-aloud Studie (N=9, Physiklehramtsstudierende und Physiklehrkräfte) wurde festgestellt, dass die Probanden mit ähnlichen Überlegungen zu den richtigen Antworten kamen wie die Experten. Ein Maß für adressatengemäßes Erklären ergibt sich bei diesem Teil des Tests aus dem Abgleich der Probandenauswahl mit der Expertenmeinung. Je höher die Übereinstimmung mit den Experten, desto besser das Ergebnis des Probanden. Die Ergebnisse aus der zweiten Stufe wurden bisher nicht bewertet sondern sollen zum Bilden von Typen verschiedener Erklärstrategien dienen.

Erste Ergebnisse aus dem Test „Impulserhaltung"
Zunächst wurde überprüft, inwieweit der Videovignettentest und der Performanztest dieselbe Fähigkeit messen. Dazu haben bislang $N = 10$ Testpersonen beide Tests absolviert. Die

Ergebnisse in beiden Testformaten korrelieren hoch ($\rho = 0.66$; $p = 0.02$). Diese Stichprobe soll auf 50 Datensätze vergrößert werden. Insgesamt wurde der Test bislang mit etwa 100 Probanden erprobt. Dabei handelte es sich um Physiklehramtsstudierende aus Deutschland und Australien, sowie Studienreferendare der Physik und Fachphysiker vom CERN (ehrenamtlich tätig als Besucherführer „Guides“). Über alle Datensätze konnte bisher lediglich eine Reliabilität von $\alpha = 0{,}41$ erreicht werden. In Tabelle 1 sind die prozentualen Ergebnisse aus dem Test zur Impulserhaltung, aufgeteilt nach sechs Kohorten jeweils für die Testteile Sachgerechtheit und Adressatengemäßheit, dargestellt. Die Unterschiede zwischen den grau und weiß hinterlegten Kohorten sind signifikant.

Sachgerechtheit

	Mittel in %
LA Uni 2 (DE)	63
Referendare	58
CERN Guides	57
LA Uni 4 (AU)	43
LA Uni 1 (DE)	40
LA Uni 3 (AU)	38

Adressatengemäßheit

	Mittel in %
LA Uni 3 (AU)	55
LA Uni 2 (DE)	53
Referendare	46
LA Uni 4 (AU)	45
LA Uni 1 (DE)	43
CERN Guides	42

Tabelle 1: Durchschn. prozentuale Ergebnisse des Tests. „LA Uni“: Lehramtsstudierende der Physik von jeweils zwei deutschen („DE“) und zwei australischen („AU“) Universitäten. „Referendare“: Studienreferendaren der Physik. „CERN Guides“: Fachwissenschaftler des CERNs, die auch als Besucherführer tätig sind.

Es fällt auf, dass zwischen den beiden deutschen Standorten erhebliche Unterschiede in beiden Testteilen bestehen. Dies kann möglicherweise daran liegen, dass die unterschiedlichen Curricula zu unterschiedlichen Lernständen führten. Die Gruppe „CERN Guides“ schneidet fachlich gut ab, hat aber offenbar Schwierigkeiten mit adressatengemäßen Aspekten des Erklärens. Beides lässt sich gut durch die Ausbildung (Fachphysik, keine didaktische Ausbildung) erklären. Eine der beiden australischen Kohorten schneidet fachlich sehr schlecht ab, nimmt beim adressatengemäßen Erklären jedoch den ersten Platz ein. Bei dieser Kohorte handelt es sich um Studierende, deren fachliche (Bachelor-)Ausbildung schon einige Jahre zurück liegt und nicht zwingend Physik als Schwerpunkt hatte. Diese Probanden waren nach ihrer Fachausbildung im Beruf tätig und haben anschließend ein weiterführendes Masterstudium zur Qualifikation als Physiklehrkraft aufgenommen, das offenbar wenig Lerngelegenheiten für physikalische Fachinhalte bietet, jedoch zu einer erfolgreichen Ausbildung mit Blick auf adressatengemäßes Erklären führt.

Ein weiterer interessanter Aspekt ist die Untersuchung von Begründungsstrategien verschiedener Kohorten hinsichtlich der Auswahl bestimmter Antworten. Die vorläufigen Ergebnisse zeigen hier einen grundsätzlichen Unterschied in der Begründung zwischen Kohorten mit und ohne Lehramtsbezug. Probanden mit einem Lehramtsstudium begründen ihr Vorgehen hochsignifikant häufiger mit schülerzentrierten Aussagen (z. B. „ich wollte es für die Schülerin nicht so kompliziert ausdrücken“). Fachphysiker begründen hingegen hochsignifikant häufiger mit inhaltlichen Überlegungen (z. B. „ich wollte ein physikalisches Konzept verdeutlichen, das in den anderen Antworten nicht vorkommt“). Keine signifikanten Unterschiede finden sich hingegen bei der Berücksichtigung der fachlichen Korrektheit (z. B. „ich wollte den Sachverhalt physikalisch so korrekt wie möglich beschreiben“).
Das hier vorgestellte Projekt ist auf drei Jahre angelegt und läuft noch bis Mitte 2018. Nächste Schritte sind die Erweiterung der Datenbasis zur Bestimmung der Korrelation zwischen den Ergebnissen aus Vignettentest und Performanztest sowie die Überarbeitung der Items zugunsten einer besseren Reliabilität.

Literatur

Aufschnaiter, C. von & Blömeke, S. (2010). Professionelle Kompetenz von (angehenden) Lehrkräften erfassen – Desiderata. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 361–367.

Bartels, H. & Kulgemeyer, C. (2016). Entwicklung eines computerbasierten Testinstruments für Erklärfähigkeit. *PhyDid B - Didaktik der Physik - Beiträge zur DPG Frühjahrstagung*.

Beck, K., Landenberger, M. & Oser, F. (Hrsg.) (2016). *Technologiebasierte Kompetenzmessung in der beruflichen Bildung; Ergebnisse aus der BMBF-Förderinitiative ASCOT:* W Bertelsmann Verlag.

Haagen-Schützenhöfer, C. & Hopf, M. (2014). Development of a two-tier test - instrument for geometrical optics. In Constantinou, C. (Hrsg.), *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research for Evidence - based Teaching and Coherent Learning*.

Hattie, J. (2009). *Visible learning*, London: Routledge.

Kulgemeyer, C. & Schecker, H. (2013). Students Explaining Science - Assessment of Science Communication Competence. *Research in Science Education*, *43*(6), 2235–2256.

Kulgemeyer, C. & Tomczyszyn, E. (2015). Physik erklären. Messung der Erklärensfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 111–126.

Merzyn, G. (2005). Junge Lehrer im Referendariat. *Der mathematische und naturwissenschaftliche Unterricht*(1), 4–7.

Miller, G. (1990). The Assessment of Clinical Skills / Competence / Performance. *Journal of the Association of American Medical Colleges*, *65*(9), 63–67.

Neuweg, G.H. (2015). Kontextualisierte Kompetenzmessung. *Zeitschrift für Pädagogik*, *61*(3), 377–383.

Rehm, M. & Bölsterli, K. (2014). Entwicklung von Unterrichtsvignetten. In Krüger, D., Parchmann, I. & Schecker, H. (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 213–225). Berlin, Heidelberg: Springer Spektrum.

Walters, K., Osborn, D. & Raven, P. (2005). The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Medical education*, *39*(3), 292–298.

Wilson, H. & Mant, J. (2011). What makes an exemplary teacher of science? The pupil's perspective. *School Science Review*, *93*(343), 121–125.

Winter, J.C.F. de, Dodou, D. & Mulder, M. (2012). Training Effectiveness of Whole Body Flight Simulator Motion. A Comprehensive Meta-Analysis. *The International Journal of Aviation Psychology*, *22*(2), 164–183.