Carina LANGE, Jena & Anke LINDMEIER, Jena

Typen von Performance Assessments im Kontext der Lehrkräftebildung in den DACH-Regionen

Theoretischer Rahmen

Die Frage nach den „besten“ Methoden zur Leitungsfeststellung von (angehenden) Lehrkräfte ist eine zentrale Frage. Messinstrumente, die Kompetenz nahe am tatsächlichen beruflichen Handeln von Lehrkräften messen, gewinnen zunehmend an Aufmerksamkeit in der Lehrkräftebildung und -forschung. Während traditionelle ‚analytische‘ Ansätze meist individuelle Einflussfaktoren (z.B. professionelles Wissen, motivationale-affektive Aspekte) prüfen, ist dieser Ansatz holistisch und zielt auf eine Prüfung von beruflicher Handlungskompetenz ab (Kaiser et al., 2017). Beispielsweise unterscheidet Lindmeier (2011) zwischen aktionsbezogener (AC; *spontane, unmittelbare Anforderungen im Unterrichtsgeschehen*) und reflexiver (RC; *Anforderungen der Unterrichtsvor-/Nachbereitung*) Kompetenz. Infolge der Konzeptualisierung, die jeweils speziell auf die Bewältigung bestimmter beruflicher Anforderungen zugeschnitten sind, wurden Bewertungsmethoden entwickelt, die sich auf diese beruflichen Anforderungen konzentrieren (Lindmeier, 2011; Shavelson, 2010). Die sogenannten Performance Assessments (PA) bewerten die Kompetenz nahe an der Leistung (Performanz) innerhalb von komplexitätsreduzierten, situierten Settings. Typisch sind beispielsweise Simulationen in denen Schüler*innenfehler diagnostiziert werden (Kron et al., 2021) oder es werden direkte Erklärungen (unter Zeitdruck) zu einer Schüler*innenfrage (Videovignette) gefordert (Jeschke et al., 2021).

Trotz vieler innovativer Ansätze gibt es noch kein gemeinsames Verständnis handlungsnaher Leistungsfeststellung. Einige domänenübergreifende internationale Reviews (z. B. Zlatkin-Troitschanskaia et al., 2016) geben einen ersten Überblick über deren Nutzung an Hochschulen – jedoch nicht spezifisch für die Lehrkräftebildung. Vorhandene PA sind meist inspiriert durch andere Forschungsfelder, wie der Psychologie oder Medizin (z. B. Einsatz standardisierter Patienten durch Schauspieler; s. Miller 1990), in denen solche PAs als Benotungsgrundlage in der Ausbildung bereits erprobt sind. Unter Berücksichtigung verschiedener Perspektiven lassen sich, neben der Authentizität, eine hohe Standardisierung und eine objektive Messung als wesentliche Gütekriterien für PA herausarbeiten (Lange & Lindmeier, submitted). Damit eignen sich PAs perspektivisch auch in besonderem Maße als alternative Prüfungsformate in der Lehrkräfteausbildung. Das Forschungsfeld ist sehr dynamisch mit hohen Innovationscharakter – ein aktueller Über-

In: IDMI-Primar Goethe-Universität Frankfurt (Hrsg.),
Beiträge zum Mathematikunterricht 2022.
*56. Jahrestagung der Gesellschaft für Didaktik der Mathematik.*WTM.
https://doi.org/10.37626/GA9783959872089.0

blick über die jüngsten Entwicklungen dieser Formate fehlt jedoch. Vor diesem Hintergrund werden in diesem Beitrag folgende Forschungsfragen untersucht: *(1) Wie unterscheiden sich die (jüngsten) performanzbasierten Messinstrumente in der DACH Lehrkräftebildung im Hinblick auf den eingesetzten Kontext, die Testmethode und deren Passung in Bezug auf die Kriterien für handlungsnahe Erhebungsmethoden? (2) Welche Typen von PA lassen sich in der DACH Lehrkräftebildung identifizieren?*

Methode

Zur Untersuchung dieser Forschungsfragen wurde im Herbst/Winter 2021 ein systematisches Literatur-Review in den Datenbanken DIE, peDOCS und Web of Science durchgeführt. Anhand festgelegter Suchparameter, sowie Ein-/Ausschlusskriterien konnten für den Zeitraum 2016-2020 insgesamt 20 Instrumente innerhalb der DACH-Regionen identifiziert werden. Eingeschlossen wurden ausschließlich Instrumente, die objektive, standardisierte Messverfahren auf der Grundlage von beobachtbarem Verhalten bezüglich beruflicher Anforderungssituationen von Lehrkräften bieten. Des Weiteren wurden nur Instrumente, die professionsspezifische Kompetenzen von Lehrkräften (z.B. Diagnose-, Unterrichtsanalysekompetenz) erheben, betrachtet – Testinstrumente, die fachspezifische Kompetenzen (z.B. Experimentierkompetenz) erheben, wurden hingegen von der Auswahl ausgeschlossen. In diesem Beitrag wird ein vorläufiger Auswertungsstand berichtet.

Ausgehend von diesen 20 Testinstrumenten wurde induktiv ein Kategoriensystem mit derzeit 14 Kategorien entwickelt, welches verschiedene Charakteristika in Bezug auf den Kontext (*Fach, Kompetenz, Einsatzzweck, Outcomes, Anzahl der Arbeitsproben*), die Testverfahren (*Stimulus und Response-Format, Materialien & Medien, Erhebungsgröße, Zeitbeschränkung, Offenheit der Aufgaben*) und die Passung der Instrumente zu den Kriterien für PA (*Situiertheit & Interaktion*) fasst. Die Codierung erfolgte mit je zwei Ratern in MaxQDA mit genügender Übereinstimmung und der resultierende Datensatz wurde deskriptiv verwendet, um die Forschungsfrage FF1 zu beantworten. Für FF2 wurde in einem ersten Schritt eine qualitative typenbildende Inhaltsanalyse (Kuckartz, 2014) durchgeführt, die in einem weiteren Schritt durch eine statistische explorative hierarchische Clusteranalyse (complete linkage) ergänzt wurde.

Ergebnisse

Die Untersuchung zeigt ein heterogenes Forschungsfeld mit Schwerpunkt in den MINT-Fächern. Trotz der großen Methodenvielfalt ist der Einsatz von Unterrichtsvignetten (*Text & Video*) in einem digitalen, offenen Antwortsetting weit verbreitet. Sowohl die qualitative typenbildende Inhaltsanalyse als

auch die statistische explorative Clusteranalyse (mittels Elbow-plot) ergaben nach derzeitigem Auswertungsstand drei Typen/Cluster von handlungsnahen Erhebungsformaten (s. Abbildung 1).

- Typ A (*action*) zeichnet sich durch hohe Handlungsanforderungen mit schnellen, unmittelbaren Reaktionen (bei Rollenspielen, Videovignetten) und einer hohen Nähe zur Realsituation aus.
- Typ B (*analyse*) bildet Formate ab, bei denen primär diagnostische bzw. analytische Tätigkeiten gefordert werden. Typischerweise werden Video- oder Textvignetten mit Unterrichtssituationen in einer computergestützten Bewertungsumgebung bereitgestellt.
- Typ C (*product*) umfasst Testinstrumente die auf die Erstellung eines Produktes (z.B. Unterrichtsentwurf) abzielen.

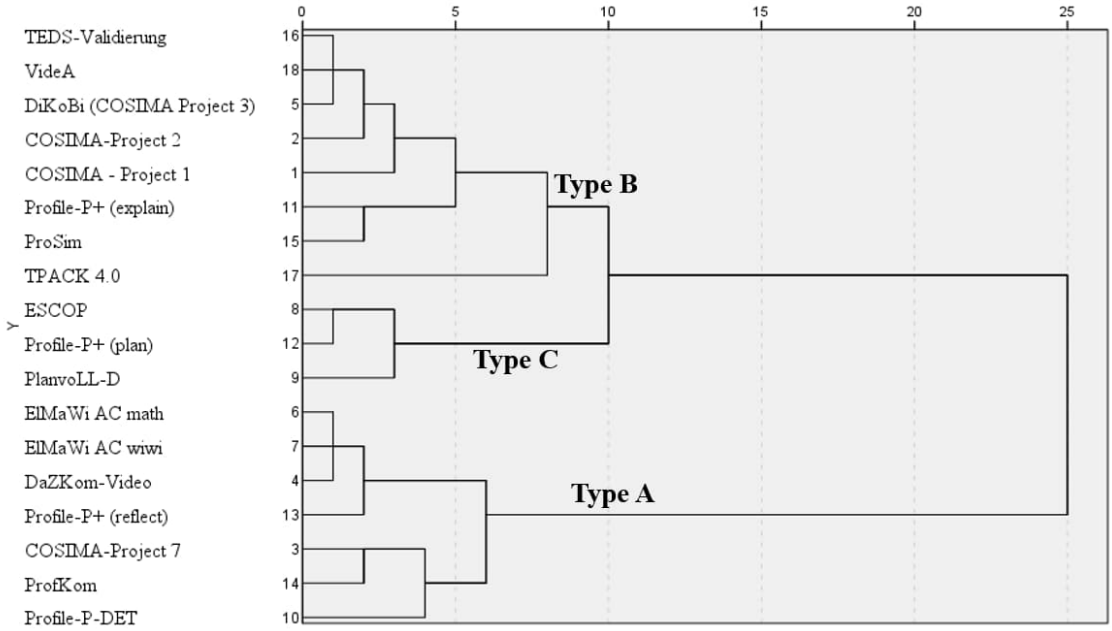

Abb. 1: Dendrogramm (explorative Clusteranalyse mit 8 Clustervariablen)

Diskussion und Ausblick

Für die Beschreibung der Typen erwies sich eine Unterteilung der Kompetenzen (*AC, RC & Mischvarianten*) nach dem vorgestellten Strukturmodell (Lindmeier, 2011) als sehr hilfreich. Betrachtet man die Übereinstimmungen der Typen mit den Kriterien für handlungsnahe Erhebungsformate, so ist diese für Typ A aufgrund der großen Ähnlichkeit zwischen der beobachteten und der interessierenden Leistungsart hoch. Gleichzeit stehen die Aspekte der Komplexität/Realitätsnähe und der Kosteneffizienz/Standardisierung in Konkurrenz zueinander (Davey et al, 2015). Einen Kompromiss bietet hier

der Typ B, der beispielsweise durch virtuelle Klassenräume oder mittels videobasierten Unterrichtsanalysen neue Wege des Messens und Prüfens aufzeigt, die über Multiple Choice Formate hinausgehen. Trotz Limitationen, wie das fehlende gemeinsame Begriffsverständnis (und damit einhergehend die hier vorgeschlagenen Kriterien) für handlungsnahe Erhebungsmethoden und der noch überschaubaren Anzahl an Instrumenten, zeigen sich erste interessante Einblicke in die Diversität des Feldes. Obwohl das Kategoriensystem sowie die Typologie noch auf ihre Robustheit hin geprüft werden müssen, zeigen sie systematisch Gestaltungsoptionen für die Entwicklung und Untersuchung neuer Mess- und Prüfungsinstrumente auf.

Literatur

Davey, T., Holland, P. W., Shavelson, R., Webb, N. M., Noreen M. & Lauress L. (2015): *Psychometric considerations for the next generation of performance assessment.* ETS.

Jeschke, C., Lindmeier, A. & Heinze, A. (2021). Vom Wissen zum Handeln: Vermittelt die Kompetenz zur Unterrichtsreflexion zwischen mathematischem Professionswissen und der Kompetenz zum Handeln im Mathematikunterricht? Eine Mediationsanalyse. *Journal für Mathematik-Didaktik, 42*(1), 159–186.

Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M. & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers—Cognitive versus situated approaches. *Educational Studies in Mathematics, 94*(2), 161–182.

Kron, S., Sommerhoff, D., Achtner, M. & Ufer, S. (2021). Selecting mathematical tasks for assessing student's understanding: Pre-service teachers' sensitivity to and adaptive use of diagnostic task potential in simulated diagnostic one-to-one interviews. *Frontiers in Education, 6,* Article 604568.

Kuckartz, U. (2014). Qualitative text analysis: A guide to methods, practice & using software. SAGE.

Lange, C. & Lindmeier, A. (submitted). *Performance assessment in the context of teacher education research – A systematical review of characteristics of instruments for an emerging topic in the DACH region.* Manuscript submitted for peer-review.

Lindmeier, A. (2011). Modeling and measuring knowledge and competences of teachers: A threefold domain-specific structure model for mathematics. Waxmann.

Miller, G. E. (1990): The assessment of clinical skills/competence/performance. In: *Academic medicine, 65*(9), S63-67.

Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training, 2*(1), 41–63.

Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M. & Lautenbach, C. (2016). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen: Ein Überblick zum nationalen und internationalen Forschungsstand.* Springer.