Jannis Zeller[1] [1]Universität Paderborn
Josef Riese[1]

Assessment des physikdidaktischen Wissens mithilfe von Machine Learning

Theoretischer Hintergrund und Forschungsfragen

Das Professionswissen von Lehrkräften steht bereits seit langem im Fokus fachdidaktischer Forschung (z. B. Shulman, 1986; Baumert & Kunter, 2006). Neben dem Fachwissen (FW) und Pädagogischem Wissen (PW) stellt das Fachdidaktische Wissen (FDW) als das Wissen, das zur Vermittlung konkreter Fachinhalte an konkrete Lernende notwendig ist, eine wichtige Domäne des Professionswissens dar. Die inneren Strukturen von PW, FW und FDW wurden bislang hauptsächlich im Rahmen normativer Modelle auf theoretischer Seite (z. B. Park & Oliver, 2008; Baumert & Kunter, 2006; Riese et al., 2017) sowie mithilfe von hierarchischen Niveaumodellen auf empirischer Seite untersucht (König, 2009; Woitkowski & Riese, 2017; Zeller et al., 2022; Schiering et al., 2023).

Das FDW wird dabei im Rahmen von (theoretischen) Strukturmodellen (u. A. zur Item-Entwicklung) typischerweise dreidimensional modelliert. Im Projekt ProfiLe-P(+) (Riese et al., 2015; Vogelsang et al., 2019) wurden dabei anhand von Literaturreviews und Expert:innenbefragungen die Dimensionen Fachinhalte, fachdidaktische Facetten (*Instruktionsstrategien*, *Schülervorstellungen*, *Experimente* und *Fachdidaktische Konzepte*) sowie kognitive Anforderungen (*Reproduzieren*, *Anwenden-Kreieren* und *Analysieren-Evaluieren*) abgebildet.

Aus empirisch-datenbasierter Perspektive stellen Zeller et al. (2024) im Rahmen einer Item-Response-basierten Analyse mithilfe des Scale-Anchoring-Verfahrens projektübergreifend fest, dass sich FDW in niedrigen Ausprägungen auf reproduktive Aspekte beschränkt, während in höheren Ausprägungen analytisch-evaluative und anwendungsorientiert-kreative Elemente hinzukommen (Datensatz s.u.). Auf diesen Ergebnissen aufbauend wurden in einem nicht-hierarchischen Ansatz mithilfe einer latenten Profilanalyse (Spurk et al., 2020) vier Kompetenzprofile bezüglich der drei o. g. kognitiven Anforderungen identifiziert (Datensatz s.u.). Diese wurden entsprechend ihrer gezeigten Stärken und Schwächen als *Low Achievers* (niedriger Score in allen Kategorien), *Applying Creatives* (Stärken im Anwenden-Kreieren, Schwächen im Analysieren-Evaluieren), *Analytic Evaluators* (Gegenteil der Applying Creatives) und *High Achievers* (hoher Score in allen Kategorien) bezeichnet (Abb. 1).

Sowohl für den Transfer der theoretischen (hier: Modellierung kognitiver Anforderungen und Facetten) und empirischen Ergebnisse (hier: latente Kompetenzprofile) in die Praxis – z. B. in Form eines Assessments zu Feedbackzwecken – als auch zur Ermöglichung weiterer skalierbarer Forschung zum FDW ist eine Automatisierung der Bepunktung des verwendeten FDW-Testinstruments (siehe z. B. Riese et al., 2015) notwendig. Da das Testinstrument größtenteils aus Aufgaben in offenem Antwortformat besteht, ist für eine solche Automatisierung die Nutzung von Methoden des Natural Language Processing (NLP) und Machine Learning (ML) naheliegend (z. B. Camus & Filighera, 2020). Zu diesem Zweck werden im vorliegenden Beitrag die folgenden Forschungsfragen untersucht:

***FF1 (Automatisches Scoring)*:** Welche Maschine-Mensch-Übereinstimmung bei der automatisierten Bepunktung erreicht ein Machine-Learning Sprachmodell (BERT, Devlin et al. 2019) auf Basis von 846 Bearbeitungen eines FDW-Testinstruments unter Nutzung eines Standard-Trainingsworkflows?

***FF2 (Kompetenzprofile)*:** Wie hoch ist die Maschine-Mensch-Übereinstimmung einer automatisierten Zuordnung von Bearbeitungen des FDW-Testinstruments zu einem prototypischen FDW-Kompetenzprofil auf Basis der Scorer-Vorhersagen?

***FF3 (Subskalen)*:** Wie ist der Zusammenhang zwischen den Vorhersagen des Scorers und der wahren Summenscores bezogen auf Subskalen des FDW (kognitive Anforderungen & Facetten)?

Dabei wurden zu FF1 neben dem hier genannten BERT-Modell auch andere Ansätze exploriert, aus Gründen der Performanz der verschiedenen Modelle und der Übersichtlichkeit dieses Beitrags wird hier aber der BERT-Ansatz fokussiert.

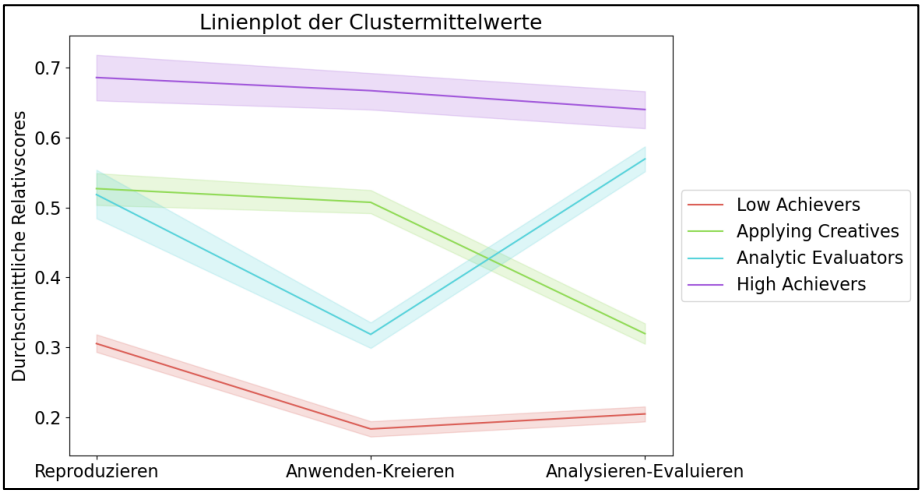

Abb. 1 Darstellung der Scores der vier Kompetenzprofile.

Stichprobe

Für die Analyse liegen 846 Bearbeitungen des FDW-Testinstruments aus dem Projekt ProfiLe-P+ vor, das in 23 offenen und 4 Multiple-Choice-Aufgaben das FDW von (angehenden) Physiklehrkräften (mittleres Studiensemester: 4,1) bezüglich des Fachinhaltes *Mechanik* sowie der o. g. Facetten und kognitiven Anforderungen erfasst. Insgesamt liegen dabei 15600 Antworten zu den offenen Aufgaben vor (454-825 pro Aufgabe), zu denen jeweils Scores (0, 1 oder 2 Punkte) durch eine trainierte Kodiererin zugeteilt wurden. Für die Bestimmung von Mensch-Mensch-Übereinstimmungs- (FF1, FF2) und Korrelationswerten (FF3), lag zudem eine Doppeltkodierung von 267 dieser Testbearbeitungen vor.

Methodik und Ergebnisse

Zu FF1 wurde das BERT-Modell (BERT-base-german-uncased) für alle Aufgaben gemeinsam für drei komplette Durchläufe der Trainingsdaten trainiert. Das Modell erreicht im Rahmen einer 10-Fold-Cross-Validation (N_{eval} = 15600) eine prozentuale Übereinstim-

mung zu den menschlichen Scores von 75,1 % bzw. ein Cohens κ von 0,560, was im Vergleich zur Mensch-Mensch-Baseline von 81,3 % bzw. κ = 0,665 als gute Übereinstimmung zu bewerten ist. Zu FF2 wurde mithilfe der Vorhersagen des BERT-Modells unter Weiternutzung der Cross-Validation-Splits ein logistisches Regressionsmodell zur Vorhersage der Kompetenzprofile trainiert und evaluiert (N_{eval} = 846), das eine prozentuale Übereinstimmung von 75,9 % bzw. κ = 0,612 erreicht. Auch dies ist im Vergleich zu einer entsprechenden Mensch-Mensch-Baseline von 73,4 % bzw. κ = 0,522, die mithilfe des wahren Cluster-Modells ermittelt wurde, als sehr gute Übereinstimmung zu werten. Zu FF3 wurden die menschlichen und maschinellen Scores gemäß der theoriebasierten Aufgabenzuordnungen zu den Summenscores bzgl. der o. g. vier Facetten und drei kognitiven Anforderungskategorien sowie zum Gesamtscore aggregiert. Die Gegenüberstellung der (menschlichen) Target-Summenscores und den entsprechenden Vorhersagen zeigt einen linearen Zusammenhang (Abb. 2). Quantitativ betrachtet korrelieren die menschlichen mit den maschinellen Summenscores im Median signifikant mit 0,86*** [0,73*** bis 0.93***], was (nicht nur) verglichen mit der Mensch-Mensch-Baseline von 0,92*** [0,79*** bis 0.96***] als hoch einzuschätzen ist.

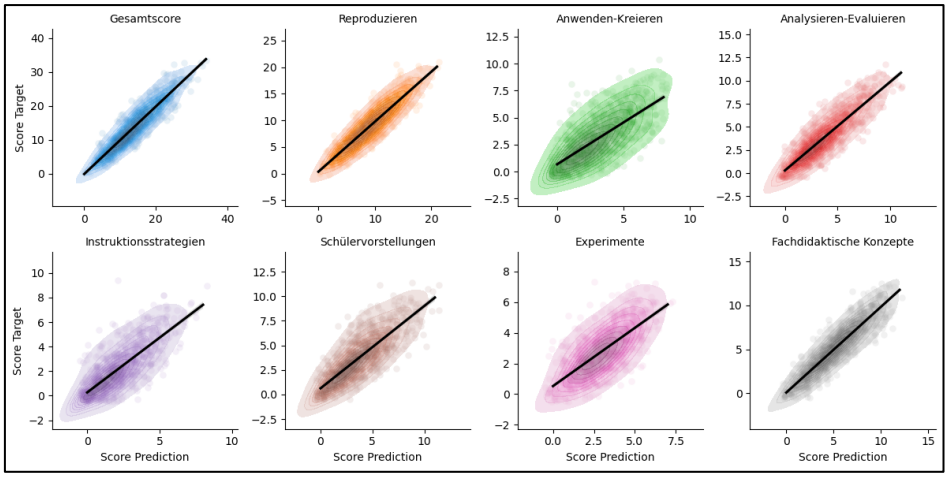

Abb. 2 Gegenüberstellung der Summenscore-Targets und -Vorhersagen

Ausblick

Insgesamt können über das BERT-Modell und die theoriebasiert abgeleiteten Aussagen aus den vorhergesagten Scores sowohl (a) typische Kompetenzprofile und (b) im Summenscores zu theoriebasierten Subskalen mit hoher Reliabilität ermittelt werden. Das Modell kann somit sowohl für die Gestaltung von Feedback, das im Rahmen der Kompetenzprofile und Subskalen auch inhaltliche Aspekte enthalten kann, als auch für potenzielle weitere FDW-Forschung eingesetzt werden, indem neue Daten skalierbar automatisiert erfasst werden. Zu diesem Zweck ist ein Webtool unter Nutzung von open-source Python-Paketen in Arbeit, das die Testbearbeitung und Auswertung enthält. Der Python-Code der explorativen Analyse der Kompetenzniveaus, der Automatisierung der Auswertung und für das Webtool wird aktuell verallgemeinert, um auch auf andere Testinstrumente und Datensätze übertragen werden zu können. Es ist geplant ihn als open-source Kooperationsprojekt anzulegen.

Literatur

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. Zeitschrift für Erziehungswissenschaft, 9(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2

Camus, L., & Filighera, A. (2020). Investigating Transformers for Automatic Short Answer Grading. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin & E. Millán (Hrsg.), Artificial Intelligence in Education (S. 43–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_8

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (Hrsg.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (S. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

König, J. (2009). Zur Bildung von Kompetenzniveaus im Pädagogischen Wissen von Lehramtsstudierenden: Terminologie und Komplexität kognitiver Bearbeitungsprozesse als Anforderungsmerkmale von Testaufgaben? Lehrerbildung auf dem Prüfstand, 2(2), 244–262. https://doi.org/10.25656/01:14703

Park, S., & Oliver, J. S. (2008). National Board Certification (NBC) as a catalyst for teachers' learning about teaching: The effects of the NBC process on candidate teachers' PCK development. Journal of Research in Science Teaching, 45(7), 812–834. https://doi.org/https://doi.org/10.1002/tea.20234

Riese, J., Gramzow, Y., & Reinhold, P. (2017). Die Messung fachdidaktischen Wissens bei Anfängern und Fortgeschrittenen im Lehramtsstudiengang Physik. Zeitschrift für Didaktik der Naturwissenschaften, 23, 99–112. https://doi.org/10.1007/s40573-017-0059-2

Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H. E., Gramzow, Y., Reinhold, P., Schecker, H., & Tomczyszyn, E. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik. Zeitschrift für Pädagogik, 61, 55–79.

Schiering, D., Sorge, S., Keller, M. M., & Neumann, K. (2023). A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK) – What constitutes high-level PCK? Journal of Research in Science Teaching, 60(1), 136–163. https://doi.org/10.1002/tea.21793

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. Educational Researcher, 15(2), 4–14. https://doi.org/10.3102/0013189X015002004

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. Journal of Vocational Behavior, 120, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Vogelsang, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., Riese, J., Schecker, H., & Schröder, J. (2019). Entwicklung von Professionswissen und Unterrichtsperformanz im Lehramtsstudium Physik–Analysen zu valider Testwertinterpretation. Zeitschrift für Pädagogik, 65(4), 473–491. https://doi.org/10.25656/01:23990

Woitkowski, D., & Riese, J. (2017). Kriterienorientierte Konstruktion eines Kompetenzniveaumodells im physikalischen Fachwissen. Zeitschrift für Didaktik der Naturwissenschaften, 23, 39–52. https://doi.org/10.1007/s40573-016-0054-z

Zeller, J., Jordans, M., & Riese, J. (2022). Ansätze zur Ermittlung von Kompetenzniveaus im Fachdidaktischen Wissen. In S. Habig & H. van Vorst (Hrsg.), Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021 (S. 768–771). Gesellschaft für Didaktik der Chemie und Physik.

Zeller, J., Schiering, D., Kulgemeyer, C., Neumann, K., Riese, J., & Sorge, S. (2024). Empirisch-kriterienorientierte Analyse des fachdidaktischen Wissens angehender Physiklehrkräfte. Welche inhaltlichen Strukturen zeigen sich über unterschiedliche Projekte hinweg? Unterrichtswissenschaft. https://doi.org/10.1007/s42010-024-00200-w